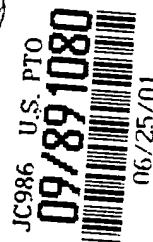


Practitioner's Docket No. JP920000066US;954-010377-US (PAR) **PATENT**

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re application of: Hurst et al. Group No.:
Serial No.: 0 /
Filed: Herewith Examiner:
For: DOCUMENT PROCESSING METHOD, SYSTEM AND MEDIUM



Assistant Commissioner for Patents
Washington, D.C. 20231

TRANSMITTAL OF CERTIFIED COPY

Attached please find the certified copy of the foreign application from which priority is claimed for this case:

Country: Japan
Application
Number: 2000-190335
Filing Date: 23.06.2000

WARNING: "When a document that is required by statute to be certified must be filed, a copy, including a photocopy or facsimile transmission of the certification is not acceptable." 37 C.F.R. 1.4(f) (emphasis added).

SIGNATURE OF PRACTITIONER

Reg. No. 29,277

David Aker
(type or print name of practitioner)

Tel. No. (203) 259-1800

Perman & Green, LLP

Customer No.: 2512

P.O. Address 425 Post Road
Fairfield, CT 06430

NOTE: The claim to priority need be in no special form and may be made by the attorney or agent, if the foreign application is referred to in the oath or declaration, as required by § 1.63.

CERTIFICATE OF MAILING (37 C.F.R. 1.8a)

I hereby certify that this correspondence is, on the date shown below is being deposited with the United States Postal Service with sufficient postage as first class mail in an envelope addressed to the Assistant Commissioner for Patents, Washington, D.C. 20231.

Date: _____

Signature _____

(type or print name of person certifying)

(Transmittal of Certified Copy [5-4])

Best Available Copy

日本国特許庁
PATENT OFFICE
JAPANESE GOVERNMENT

JC986 U.S. PTO
09/891080
06/25/01

別紙添付の書類に記載されている事項は下記の出願書類に記載されて
いる事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed
with this Office.

出願年月日
Date of Application:

2000年 6月23日

出願番号
Application Number:

特願2000-190335

出願人
Applicant(s):

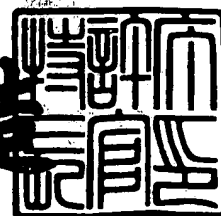
インターナショナル・ビジネス・マシーンズ・コーポレー
ション

CERTIFIED COPY OF
PRIORITY DOCUMENT

2000年12月22日

特許庁長官
Commissioner,
Patent Office

及川耕造



出証番号 出証特2000-3107287

【書類名】 特許願

【整理番号】 JP9000066

【提出日】 平成12年 6月23日

【あて先】 特許庁長官殿

【国際特許分類】 G06F 17/21

【発明者】

【住所又は居所】 神奈川県大和市下鶴間 1 6 2 3 番地 1 4 日本アイ・ビー・エム株式会社 東京基礎研究所内

【氏名】 マシュー・フランシス・ハースト

【発明者】

【住所又は居所】 神奈川県大和市下鶴間 1 6 2 3 番地 1 4 日本アイ・ビー・エム株式会社 東京基礎研究所内

【氏名】 那須川 哲哉

【特許出願人】

【識別番号】 390009531

【氏名又は名称】 インターナショナル・ビジネス・マシーンズ・コーポレーション

【代理人】

【識別番号】 100086243

【弁理士】

【氏名又は名称】 坂口 博

【復代理人】

【識別番号】 100112520

【弁理士】

【氏名又は名称】 林 茂則

【電話番号】 046-277-0540

【選任した代理人】

【識別番号】 100091568

【弁理士】

【氏名又は名称】 市位 嘉宏

【選任した復代理人】

【識別番号】 100110607

【弁理士】

【氏名又は名称】 間山 進也

【選任した復代理人】

【識別番号】 100098121

【弁理士】

【氏名又は名称】 間山 世津子

【手数料の表示】

【予納台帳番号】 091156

【納付金額】 21,000円

【提出物件の目録】

【物件名】 明細書 1

【物件名】 図面 1

【物件名】 要約書 1

【包括委任状番号】 9706050

【包括委任状番号】 9704733

【プルーフの要否】 要

【書類名】 明細書

【発明の名称】 文書処理方法、文書処理システムおよび媒体

【特許請求の範囲】

【請求項 1】 文書処理システムにおいて空白文字またはタブその他のスペースで文字がレイアウトされている文書から意味のあるテキストブロックを抽出する文書処理方法であって、

前記文書から文字、記号、空白文字その他シンボルで構成されるオブジェクトを生成するステップと、

前記オブジェクト間の接続候補を生成するステップと、

前記接続候補の接続の妥当性を言語モデルを用いて評価するステップと、

を含む文書処理方法。

【請求項 2】 前記接続候補の接続が妥当であると判断された場合には、前記接続候補の接続元および接続先の前記オブジェクトを結合するステップをさらに含む請求項 1 記載の文書処理方法。

【請求項 3】 前記オブジェクトは、前記文書の空間位置を示す座標に関連付けられて生成される請求項 1 または 2 記載の文書処理方法。

【請求項 4】 前記オブジェクトの結合により前記テキストブロックが生成され、前記テキストブロックは、前記オブジェクトを含む最小面積の方形領域で定義され、前記文書における前記方形領域の対角 2 点の座標で位置が特定される請求項 3 記載の文書処理方法。

【請求項 5】 前記オブジェクト間の接続候補は、接続元オブジェクトの右方に隣接するオブジェクトとの接続、または、前記接続元オブジェクトが存在する行の次行に存在し、前記接続元オブジェクトより左方に位置する何れかのオブジェクトとの接続である請求項 1 ～ 4 の何れか一項に記載の文書処理方法。

【請求項 6】 前記言語モデルは n グラムモデルである請求項 1 ～ 5 の何れか一項に記載の文書処理方法。

【請求項 7】 前記オブジェクトの生成ステップには、
前記文書の空間座標に関連付けて、前記座標毎のシンボルを取得するステップと、

前記文書の 1 行内の前記シンボルのタイプを判断し、1 つまたは連続する文字、記号その他のキャラクタで構成されるトークン、または、1 つまたは連続する空白文字で構成されるスペースを生成するステップと、

前記スペースの上下方向の隣接関係を判断し、複数行にわたるスペースで構成されるストリームを生成するステップと、

前記トークンと前記ストリームとの位置関係を判断し、前記トークンまたは空白文字を含む初期テキストブロックを生成するステップと、

を含む請求項 1 ～ 6 の何れか一項に記載の文書処理方法。

【請求項 8】 前記トークンまたはスペースを生成するステップにおいて、前記タイプが空白文字でないと判断され、前記行内において隣接するシンボルの前記タイプが同じであると判断された場合には、前記シンボルを連続するキャラクタで構成される 1 つのトークンとして記録し、前記タイプが空白文字であると判断された場合には前記シンボルを 1 つまたは連続するスペースとして記録し、

前記ストリームを生成するステップにおいて、前記スペースが異なる行において上または下の方向に隣接すると判断された時には前記スペースをストリームとして記録し、

前記オブジェクトの生成ステップにおいて、1 行内の 2 つのトークンに挟まれたスペースがストリームでない場合には、前記 2 つのトークンとその間のスペースを初期テキストブロックとして結合する請求項 7 記載の文書処理方法。

【請求項 9】 前記初期テキストブロックとその接続候補の全てとを生成するステップと、

前記初期テキストブロックおよび接続候補の全てから単一要素の初期テキストブロックおよび接続候補を抽出するステップと、

前記単一要素の初期テキストブロックおよび接続候補の接続妥当性を言語モデルを用いて判断するステップと、

前記接続妥当性が妥当であると判断された時には、前記単一要素の初期テキストブロックを結合するステップと、

を含む請求項 7 または 8 記載の文書処理方法。

【請求項 10】 前記オブジェクト、初期テキストブロックまたはそれらが

結合されたテキストブロックとの間に単一の接続候補のみが存在する時には、言語モデルを用いた接続妥当性を判断することなくこれらを結合する請求項 1 ～ 9 の何れか一項に記載の文書処理方法。

【請求項 1 1】 空白文字またはタブその他のスペースで文字がレイアウトされている文書から意味のあるテキストブロックを抽出する文書処理システムであって、

前記文書から文字、記号、空白文字その他シンボルで構成されるオブジェクトを生成する手段と、

前記オブジェクト間の接続候補を生成する手段と、

前記接続候補の接続の妥当性を言語モデルを用いて評価する手段と、

を含む文書処理システム。

【請求項 1 2】 前記接続候補の接続が妥当であると判断された場合には、前記接続候補の接続元および接続先の前記オブジェクトを結合する手段をさらに含む請求項 1 1 記載の文書処理システム。

【請求項 1 3】 前記オブジェクトは、前記文書の空間位置を示す座標に関連付けられて生成される請求項 1 1 または 1 2 記載の文書処理システム。

【請求項 1 4】 前記オブジェクトの結合により前記テキストブロックが生成され、前記テキストブロックは、前記オブジェクトを含む最小面積の方形領域で定義され、前記文書における前記方形領域の対角 2 点の座標で位置が特定される請求項 1 3 記載の文書処理システム。

【請求項 1 5】 前記オブジェクト間の接続候補は、接続元オブジェクトの右方に隣接するオブジェクトとの接続、または、前記接続元オブジェクトが存在する行の次行に存在し、前記接続元オブジェクトより左方に位置する何れかのオブジェクトとの接続である請求項 1 1 ～ 1 4 の何れか一項に記載の文書処理システム。

【請求項 1 6】 前記言語モデルは n グラムモデルである請求項 1 1 ～ 1 5 の何れか一項に記載の文書処理システム。

【請求項 1 7】 前記オブジェクトの生成手段には、

前記文書の空間座標に関連付けて、前記座標毎のシンボルを取得する手段と、

前記文書の 1 行内の前記シンボルのタイプを判断し、1 つまたは連続する文字、記号その他のキャラクタで構成されるトークン、または、1 つまたは連続する空白文字で構成されるスペースを生成する手段と、

前記スペースの上下方向の隣接関係を判断し、複数行にわたるスペースで構成されるストリームを生成する手段と、

前記トークンと前記ストリームとの位置関係を判断し、前記トークンおよび空白文字を含む初期テキストブロックを生成する手段と、

を含む請求項 1 1 ～ 1 6 の何れか一項に記載の文書処理システム。

【請求項 1 8】 前記トークンまたはスペースを生成する手段において、前記タイプが空白文字でないと判断され、前記行内において隣接するシンボルの前記タイプが同じであると判断された場合には、前記シンボルを連続するキャラクタで構成される 1 つのトークンとして記録し、前記タイプが空白文字であると判断された場合には前記シンボルを 1 つまたは連続するスペースとして記録し、

前記ストリームを生成する手段において、前記スペースが異なる行において上または下の方向に隣接すると判断された時には前記スペースをストリームとして記録し、

前記オブジェクトの生成手段において、1 行内の 2 つのトークンに挟まれたスペースがストリームでない場合には、前記 2 つのトークンとその間のスペースを初期テキストブロックとして結合する請求項 1 7 記載の文書処理システム。

【請求項 1 9】 前記初期テキストブロックとその接続候補の全てとを生成する手段と、

前記初期テキストブロックおよび接続候補の全てから単一要素の初期テキストブロックおよび接続候補を抽出する手段と、

前記単一要素の初期テキストブロックおよび接続候補の接続妥当性を言語モデルを用いて判断する手段と、

前記接続妥当性が妥当であると判断された時には、前記単一要素の初期テキストブロックを結合する手段と、

を含む請求項 1 7 および 1 8 記載の文書処理システム。

【請求項 2 0】 前記オブジェクト、初期テキストブロックまたはそれらが

結合されたテキストブロックとの間に単一の接続候補のみが存在する時には、言語モデルを用いた接続妥当性を判断することなくこれらを結合する請求項 1 1 ～ 1 9 の何れか一項に記載の文書処理システム。

【請求項 2 1】 空白文字またはタブその他のスペースで文字がレイアウトされている文書から意味のあるテキストブロックを抽出するプログラムが記録されたコンピュータ可読な記録媒体であって、前記プログラムは、

前記文書から文字、記号、空白文字その他のシンボルで構成されるオブジェクトを生成し、

前記オブジェクト間の接続候補を生成し、

前記接続候補の接続の妥当性を言語モデルを用いて評価し、

前記接続候補の接続が妥当であると判断された場合には、前記接続候補の接続元および接続先の前記オブジェクトを結合する手順をコンピュータに実行させるものである記録媒体。

【発明の詳細な説明】

【 0 0 0 1 】

【発明の属する技術分野】

本発明は、多段組、表、箇条書き、タイトル付け等任意にレイアウトされた文書から意味のあるテキストブロックを抽出する文書処理方法、システムおよび媒体に関する。本発明の技術はテキストマイニング処理、機械翻訳等自然言語文書処理の前処理に適用し得る。

【 0 0 0 2 】

【従来の技術】

近年コンピュータネットワーク上で流通する電子化された文書あるいはスキャナ等の読み取り装置で電子化された文書が膨大に蓄積されており、これら文書の活用が望まれている。蓄積されている文書の活用方法の 1 つにテキストマイニング処理（文書の概要を自動的に把握し、内容の経時的変化や傾向を把握等する文書検索処理の一種）がある。また、機械翻訳の元データとして活用される場合がある。

【 0 0 0 3 】

これら蓄積文書の活用を考慮すれば、文書のレイアウトを解析する必要がある。一般に流通している文書たとえばホームページにアップロードされる文書等では人間が視覚により把握しやすいようにレイアウトされている。また、スキャナ等により電子化された文書データでは、原稿は紙媒体の文書であり、通常の印刷様式に基づいてレイアウトされている。これらレイアウトされた文書には、文章の本体である段落のほかに、タイトル、ヘッダ、リスト、表等が含まれ、また段落も2段組等多段で表示される場合が多い。さらに表の中には、横書きの要素ばかりでなく、縦書き要素が含まれる場合もある。このため、元文書のレイアウトを考慮しなければ満足な文書解析を自動的に行うことは困難である。

【0004】

レイアウト解析の方法には、空間的な特徴に着目する方法がある。たとえば空白に着目し、空白行が挿入されている場合にはその後段は段落であると推定できる。

【0005】

【発明が解決しようとする課題】

ところが、これら空間的な特徴により意味のあるテキストブロックを抽出するには限界がある。たとえば段落要素(文章がページの纏まった領域でタイプされているようなテキスト文書)の場合と表中のテキストの場合を比較すれば、各々空白の使い方が相違する。つまり行頭に空白文字(またはタブによる空白)が表示されている時には段落の始めであることが認められるが、表中の空白は通常そのようには配置されない。また、箇条書き等リスト表示される時には行頭にインデントが付されたり、行間に空白行が挿入される。これら多様にレイアウトされたテキスト文書を一元的に空白の有無のみで解析するのは困難である。

【0006】

また、仮にレイアウトからテキストのブロックが抽出されても、そのブロック内の文章(あるいは単語の連なり)の意味上の評価が行われているわけではない。このため、特に表や見出し、リスト等段要素のように纏まったテキスト文書として表示されていない要素の場合にはブロックが分断され、その意味が正確に読み取れない。

【 0 0 0 7 】

ところで、蓄積されている文書の高度な利用（たとえばテキストマイニング）の場合には、文書の内容を自動的に判別する必要があるが、内容的に重要なメッセージは段落要素よりも表、リスト（箇条書き）等に含まれることが多い。従来、空間的な特徴に基づくレイアウト解析の場合には、その解析の困難性から表、リスト（箇条書き）等の要素の解析を断念していた（あるいは要素が断片化されるため、その後の利用が困難であった）。しかしながら、むしろ後の高度利用を考慮すれば、これら表、リスト（箇条書き）等の要素にこそ重要なメッセージが内包されている可能性が高く、後の意味解析にまで適用し得る形態で抽出することが望まれる。

【 0 0 0 8 】

本発明の目的は、表、箇条書き、多段組等任意にレイアウトされた文書から意味のあるテキストブロックを抽出する技術を提供することにある。

【 0 0 0 9 】

【課題を解決するための手段】

本願の発明の概略を説明すれば、以下の通りである。すなわち、本発明は、空白文字等のスペースにより任意にレイアウトされた文書から、たとえば単語に代表されるトークン、1つまたは連続した空白文字からなるスペース、またはこれらの結合等文書を構成するオブジェクトを生成する。オブジェクトは文書の空間位置に関連付けて生成する。そしてオブジェクト間の接続候補を生成する。オブジェクトと接続候補はグラフ理論の点（ノード）と辺（弧）に対応付けることができる。各リンクの妥当性を言語モデル（たとえばNグラムモデル）により判断し、接続候補（リンク）が妥当であると判断されればオブジェクトを結合する。

【 0 0 1 0 】

このように文書进行处理することにより、様々にレイアウトされた文書において意味のあるテキストブロックを効率的に抽出することが可能になる。

【 0 0 1 1 】

オブジェクト間に生成される接続候補（リンク）は、オブジェクトの右側（横書き文書の場合）の他のオブジェクトあるいは次行（横書き文書の場合）のそれ

より左側に位置するオブジェクトとの間に生成できる。

【 0 0 1 2 】

オブジェクトの生成にあたっては、1行内の複数トークンあるいはスペースを結合して初期テキストブロックを生成できる。この時、列方向に連続するスペースをストリームとして定義し、このストリームの文書内空間位置の情報を利用できる。つまり、ストリームで分断されるトークンは結合されず、ストリームで分断されないトークンは結合される。

【 0 0 1 3 】

トークンの生成は、文書内空間座標に対応付けて1文字ずつキャラクタを取得し、このキャラクタのタイプ（アルファベット文字であるか、記号であるか、空白文字であるか等）を判断できる。タイプが同じキャラクタが連続する場合には1つのトークンとして記録できる。空白が連続する場合にはスペースとして記録できる。

【 0 0 1 4 】

オブジェクト間の接続妥当性は、初期テキストブロックの接続妥当性で評価できる。また、この評価は複数段階行える。まず、複数のオブジェクト間の全ての接続候補（接続可能性）において、単一要素のサブクラスタ（単一の入出次数を有するオブジェクトの集合）を生成できる。そして、この単一要素サブクラスタにおける接続妥当性を言語モデルを用いて評価できる。接続が妥当であれば、この単一要素サブクラスタを構成するオブジェクト（初期テキストブロック）を結合（マージ）できる。次に、マージした後のテキストブロック間の接続妥当性を同様の手法により評価できる。このようにして効率的に全ての接続候補を評価することができる。なお、接続候補が単一の場合には、言語モデルを用いた評価を行うことなくその接続候補の接続元および接続先のオブジェクト（初期テキストブロック、マージしたテキストブロック）を結合できる。

【 0 0 1 5 】

【発明の実施の形態】

以下、本発明の実施の形態を図面に基づいて詳細に説明する。ただし、本発明は多くの異なる態様で実施することが可能であり、本実施の形態の記載内容に限

定して解釈すべきではない。なお、実施の形態の全体を通して同じ要素には同じ番号を付するものとする。

【 0 0 1 6 】

以下の実施の形態では、主に方法またはシステムについて説明するが、当業者であれば明らかなとおり、本発明は方法、システムその他、コンピュータで使用可能なプログラムが記録された媒体としても実施できる。したがって、本発明は、ハードウェアとしての実施形態、ソフトウェアとしての実施形態またはソフトウェアとハードウェアとの組合せの実施形態をとることができる。プログラムが記録された媒体としては、ハードディスク、CD-ROM、光記憶装置または磁気記憶装置を含む任意のコンピュータ可読媒体を例示できる。

【 0 0 1 7 】

本発明の文書処理方法は、一般的なコンピュータシステムを用いて実現できる。本発明のシステムは、スタンドアロンのコンピュータシステムまたは複数のコンピュータシステムで構成されたコンピュータネットワークで実現できる。図1 (a) はスタンドアロンのコンピュータを構成の概略を示し、図1 (b) は、コンピュータネットワークの場合を示す。

【 0 0 1 8 】

コンピュータシステムには、中央演算処理装置1 (CPU)、主記憶装置2 (メインメモリ: RAM)、不揮発性記憶装置3 (ROM) 等を有し、バス4 で相互に接続される。バス4 には、その他コプロセッサ、画像アクセラレータ、キャッシュメモリ、入出力制御装置 (I/O) 等が接続されてもよい。また、バス4 には、適当なインターフェイスを介して外部記憶装置5、データ入力デバイス6、表示デバイス7、通信制御装置8 等が接続される。その他、一般的にコンピュータシステムに備えられるハードウェア資源を備えることが可能なことは言うまでもない。

【 0 0 1 9 】

外部記憶装置5 は代表的にはハードディスク装置が例示できるが、これに限られず、光磁気記憶装置、光記憶装置、フラッシュメモリ等半導体記憶装置も含まれる。なお、データの読み出しのみに利用できるCD-ROM等の読み出し専用

記憶装置もデータあるいはプログラムの読み出しに適用する場合には外部記憶装置に含まれる。

【 0 0 2 0 】

データ入力デバイス 6 には、キーボード等の入力装置、マウス 9 等ポインティングデバイスを備えることができる。データ入力デバイスにはスキャナ等の画像読み取り装置、音声入力装置も含む。表示装置 7 としては、C R T、液晶表示装置、プラズマ表示装置が例示できる。

【 0 0 2 1 】

複数のコンピュータシステムで本発明を実現する場合、図 1 (b) に示すように、各コンピュータシステムは、L A N、W A N 等で接続されていてもよく、また、インターネットを介して接続されても良い。これら接続に用いられる通信回線は、専用線、公衆回線の何れでも良い。コンピュータシステムには、パーソナルコンピュータ 1 0、ワークステーション 1 1、メインフレームコンピュータ 1 2 等各種のコンピュータが含まれる。

【 0 0 2 2 】

コンピュータシステムが複数接続されたコンピュータネットワークにおいては、一部のプログラムをユーザのコンピュータで、一部のプログラムをリモートコンピュータで分散的に処理を実行できる。また、プログラムで利用されるデータは、それがどのコンピュータに記録されているかは問われない。つまり、データの所在に関する情報（アドレス）が明らかである限り、データあるいはプログラムの格納場所はコンピュータネットワーク上の任意の場所とすることができる。各ネットワークコンピュータ間の通信には公知の通信技術を適用でき、たとえば T C P / I P、H T T P 等のプロトコルを用いることができる。また、各記憶装置に記録された各ファイル（データあるいはプログラム）の存在箇所（アドレス）は、D N S、U R L 等を用いて特定できる。なお、本明細書においてインターネットという用語には、イントラネットおよびエクストラネットも含むものとする。インターネットへのアクセスという場合、イントラネットやエクストラネットへのアクセスをも意味する。コンピュータネットワークという用語には、公的にアクセス可能なコンピュータネットワークと私的なアクセスしか許可されない

コンピュータネットワークとの両方が含まれるものとする。

【 0 0 2 3 】

次に、本明細書で用いる用語を説明する。特に言及した場合を除き、本明細書では以下の概念で用語を用いる。

【 0 0 2 4 】

「オブジェクト」は、以下に説明するキャラクタ、スペース、トークン、テキストブロック等文書を構成する要素を総称する。

【 0 0 2 5 】

「シンボル」とは、空白文字を含むキャラクタのセットであり、「キャラクタ」とは、a,b,c等のアルファベット文字、記号等の独立したシンボルセットをいう。図3に示す網掛けの部分21の「N」「S」「H」「R」「R」はキャラクタの例である。なお、漢字等の2バイト文字もキャラクタに含む。

【 0 0 2 6 】

「スペース」とは1行内の空白文字あるいはその連続したものをいう。図3に示す網掛けの部分22はスペースの例である。2バイトの空白文字も含むが、1バイト空白文字の2文字分の連続と等価である。

【 0 0 2 7 】

「トークン」とは同一行内のキャラクタまたはその連続したものをいう。図3に示す網掛けの部分23の「Exercise」はトークンの例である。

【 0 0 2 8 】

「テキストブロック」とはトークンのセットである。テキストブロックはトークンが含まれる最小面積の方形で表され、左上及び右下の座標で文書中の位置が記述できる。図4に示す網掛け部分24はテキストブロックの例であり、9個のトークン「Number」「of」「Securities」「Underlying」「Options」「Granted」「(」「#」「)」が含まれる。なお、テキストブロックにはスペースが含まれてもよい。

【 0 0 2 9 】

後に説明するようにキャラクタおよびスペースはトークンの生成に用いられ、トークンはテキストブロックの生成に用いられる。トークン、スペースおよびテ

キストブロックは、その位置座標と共にデータベースに記録され蓄積される。このように位置座標と共にトークン、スペースおよびテキストブロック（オブジェクト）をデータベースに記録するため、これらオブジェクトの文書における実際の位置検索が速やかにできるようになる。

【 0 0 3 0 】

また、本明細書ではオブジェクトの抽象化のために「グラフ」、「グラフセット」、「単一要素サブクラスタ」および「複雑度」の概念を用いる。

【 0 0 3 1 】

「グラフ」とはノード（点）と弧（辺）のセットである。図 5 (a) にグラフの一例を示す。ノード 2 5 間は方向を持つ弧 2 6 で接続される。弧 2 6 の始点はソースであり弧の終点はシンクである。

【 0 0 3 2 】

「グラフセット」とは、グラフのセットである。図 5 (b) にグラフセットの一例を示す。

【 0 0 3 3 】

「単一要素サブクラスタ」とは、グラフの部分グラフであり、各々のノードから出る弧の数（出次数）およびノードに入る弧の数（入次数）が 1 のグラフである。図 6 に単一要素サブクラスタの一例を示す。矢印の左側に示すグラフから単一要素サブクラスタを抽出したものが矢印の右側に示されている。2 つのノード 2 7, 2 8 については入次数が 1 であるが、そのソースであるノード 2 9 の出次数が 2 であるからノード 2 7, 2 8 が除外されて単一要素サブクラスタが構成される。

【 0 0 3 4 】

グラフセットあるいはグラフを構成するノードに関連するリンク（ノードに入出する弧数）の度合いは複雑度で表される。「複雑度」とは、ソース（ノード）から出るリンク（弧）の数とシンク（ノード）に入るリンク（弧）の数の和である。従って、単一要素サブクラスタの複雑度は 2 となる。また、あるグラフセットにおける最大複雑度は、グラフセット内の全ての弧における最大複雑度である。

【 0 0 3 5 】

本発明では、これらグラフの概念を用いて文書を表現する。各ノードがテキストブロックに対応し、弧がテキストブロック間のリンク（接続関係）に対応する。シンク（弧の終点となるノード）はソース（弧の始点となるノード）からの接続可能性のあるテキストブロックを表す。単一のテキストブロックから複数の弧が出ている場合には、複数のテキストブロックへの接続可能性を有することになる。

【 0 0 3 6 】

また、本発明ではテキストブロックと同様にスペースをグループ化してストリームを生成する。「ストリーム」とは、文書内の各行において上下に位置する各ノードを相互に接続したスペース（ノード）で構成されるグラフである。ストリームの長さは上下に延びる空白行の行数で表される。図 7 にストリームの一例を示す。図示するようにスペース 3 0 の上下の広がりにより長さ 5 のストリームが構成されている。

【 0 0 3 7 】

以下、本実施の形態の文書処理方法を説明する。まず、前記したようなシステムに処理対象となる文書を入力する。入力とはたとえばスキャナ等の読み取り装置で入力されるほか、既に電子化された文書データとして入力される。ただし、電子化された文書であっても、空白文字、タブ等でレイアウトされている文書であれば十分であり、高度に構造化されている必要はない。図 2 は、本システムに入力される文書の一例を示す図である。図 2 に示す文書は文字等のキャラクタで構成された単一ファイルである。ここで、文書とは、一対の座標で各々独立に特定される文字、空白等シンボルの集合と定義できる。図 2 において、左上の位置を（0，0）とし、横方向（x 方向）に 1 文字ずつ x 座標指標が増加し、下方向（y 方向）に 1 行ずつ y 指標が増加するように座標を定義付ける。たとえば上から 6 行目の左側に表示されているテキスト「Name」の「N」の座標は（2，5）である。また、文書の行数は y の最大値 maxy であり、行内におけるシンボルの数は x の最大値 maxx である。このように座標に関連付けて 1 文字ずつデータベースに記録する。なお、次に説明するトークン生成処理と連動して、各行ごとに行の初

めから 1 文字ずつ右方向にシンボルを読み取る方式により入力されてもよい。

【 0 0 3 8 】

図 8 は本実施の形態の処理の概要を示したフローチャートである。文書データを入力後、初期化処理を行い（ステップ 3 1）、次に単一要素サブクラスタの結合処理を行う（ステップ 3 2）。その後、ユニークなリンクを有するクラスタ間を結合し（ステップ 3 3）、最後に残ったクラスタのリンクを評価してテキストブロックを生成する（ステップ 3 4）。

【 0 0 3 9 】

初期化処理を説明する。初期化処理は、4 つのステップで行われる。第 1 のステップはトークン生成ステップである。第 2 のステップはストリーム生成ステップである。第 3 のステップは初期テキストブロックの生成ステップであり、第 4 のステップは初期リンクの生成ステップである。

【 0 0 4 0 】

図 9 は、トークン生成ステップの一例を示したフローチャートである。ステップ 4 0 から処理を開始する。文書データの列方向の指標 y に 0 を代入して初期化し（ステップ 4 1）、 y が最大行数 $max\ y$ より小さいかを判断する（ステップ 4 2）。ステップ 4 2 の判断が no なら処理を終了し（ステップ 4 3）、それ以外は以下の処理を行う。なお、図中「=」の記号は代入記号であり、以下同様である。

【 0 0 4 1 】

文書データの行方向の指標 x に 0 を代入して x を初期化し（ステップ 4 4）、変数 $start$ に x を代入する（ステップ 4 5）。 x が最大文字数 $max\ x$ を超えないかを判断し（ステップ 4 6）、ステップ 4 6 の判断が no なら y を 1 つ増加して（ステップ 4 7）ステップ 4 2 に戻り、次行の処理に進む。それ以外の場合には以下の処理（ y 行内のキャラクタのトークン化）を行う。

【 0 0 4 2 】

まず、変数 T に関数 $char_type(x, y)$ の戻り値を代入する（ステップ 4 8）。関数 $char_type(x, y)$ は、座標 (x, y) の位置にあるシンボルのキャラクタタイプを戻り値として返す関数である。本実施の形態では、アルファベット、数字、句

読点、スペースをキャラクタタイプとして考慮する。ただし、日本語等英語以外の言語を処理する時には漢字等他の文字を考慮してもよいことは勿論である。

【 0 0 4 3 】

次に、`char_type(x,y)`の戻り値と変数`T`の値が等しいかを判断する（ステップ49）。なお、図中「==」の記号は両辺の値が等しいか否かを判断する記号であり、以下同様である。最初のループでは前記判断は「真（y e s）」になるのでステップ50に進み`x`を1つ増加する。`x`が`max x`以下であることを判断し（ステップ51）、`y e s`であればステップ49に戻る。`x`が1つ増加するのでステップ49では`y`行内の隣接するシンボルのタイプを検査することになる。シンボルタイプが同じ（ステップ49の判断が`y e s`）の場合には`x`が`max x`を超えない範囲でステップ50、51のループを繰り返し、タイプの異なるシンボルが検出されるまでこのループが繰り返される。異なるタイプのシンボルが検出されると（ステップ49の判断が`n o`）ステップ52に進み、変数`token`にこれら同一タイプの連続するシンボルの座標が入力される（ステップ52）。なお、ステップ51で`n o`と判断された場合（行末まで処理が進んだ場合）にもステップ52に進む（ステップ51）。

【 0 0 4 4 】

次にキャラクタタイプがスペースであるかを判断する（ステップ53）。スペースであるときには変数`token`をスペースデータベースに追記し（ステップ54）、スペースでない場合には変数`token`をトークンデータベースに追記する（ステップ55）。その後ステップ45に進み、前記処理を繰り返す。

【 0 0 4 5 】

このようにして入力文書のトークン化処理が行われる。なお、前記の通りトークン化と同時にスペースの検出も行われる。図10は、トークン化処理が終了した後の文書の一例を示す。たとえば0行目に着目すれば、`x`が0～25の範囲でスペースが検出され、1つのスペース（`token`）としてスペースデータベースに記録される。`x`が26～35の範囲の「i」「n」「d」「i」「v」「i」「d」「u」「a」「l」が同一タイプのキャラクタなのでトークン「i n d i v i d u a l」が生成され、トークンデータベースに記録される。なお、トークン

生成の手法としてchar_type関数を用いる例を示したが、その他個別キャラクタからトークンを発生させる方法は種種存在し、上記の手法には限られない。

【0046】

次に、ストリーム生成手法を説明する。ストリームはスペースデータベースを用いて計算される。図11は、ストリームサイズの計算方法の一例を示したフローチャートである。文書データの列方向の指標yに0を代入して初期化し（ステップ56）、yが最大行数max yより小さいかを判断する（ステップ57）。ステップ57の判断がnoなら処理を終了し（ステップ58）、それ以外は以下の処理を行う。

【0047】

y行に存在するスペースを変数Sに代入し（ステップ59）、変数nにSの数|S|を代入する（ステップ60）。なお、|O|はオブジェクトOの数を示すスカラー値であり、以下同様である。また、変数Sはベクトル量であり、ベクトルの各要素にスペース（オブジェクト）が代入される。以下変数A、変数Tにおいて同様である。

【0048】

変数iに0を代入して初期化し（ステップ61）、iがnより小さいかを判断する（ステップ62）。ステップ62の判断がyesなら変数spaceにi番目のスペースS[i]を代入し（ステップ63）、変数Aにy-1行目におけるスペースのうち、スペースS[i]のx方向位置が一致するスペースを代入する（ステップ64）。そして変数space.aboveに、スペースAのうち何れかのスペースs'の持つ上部スペース数の最大値(max s'.above)に1を加えた数を代入する（ステップ65）。ここで、変数s.aboveには、スペースsの上部にあるスペース数が記録されている。つまり前記処理により、スペースspaceの上部に存在するスペース数として、スペースspaceにx方向位置が一致するベクトル量Aの要素s'のうち最大の上部スペース数(max s'.above)に1を加えた数が代入される。space.aboveは、spaceの上部に存在する連続したスペースの数（行数）を示す。

【0049】

その後、iに1を加えて（ステップ66）、ステップ62に戻る。このように

して y 行目に存在する各スペース (S) の上部にあるスペース数が計算される。
この操作を $max y$ まで繰り返す (ステップ 67)。

【 0 0 5 0 】

上記手段により任意のスペース上にあるスペース数が計算でき、所定の閾値を超えた時にはこれをストリームと判断してストリームデータベースに記録できる。図 12 は、ストリームを計算した後の結果を示す文書である。網掛けブロックで示したスペース 68 がストリームを構成する。

【 0 0 5 1 】

次に、初期テキストブロックの生成方法を説明する。初期テキストブロックは、トークンデータベース、スペースデータベースおよびストリームデータベースを用いて計算される。図 13 は初期テキストブロックの計算方法の一例を示したフローチャートである。文書データの列方向の指標 y に 0 を代入して初期化し (ステップ 69)、 y が最大行数 $max y$ より小さいかを判断する (ステップ 70)。ステップ 70 の判断が no なら処理を終了し (ステップ 71)、それ以外は以下の処理を行う。

【 0 0 5 2 】

y 行に存在するトークンを変数 T に代入し (ステップ 72)、 y 行に存在するスペースを変数 S に代入する (ステップ 73)。そして変数 n に T の数 $|T|$ を代入する (ステップ 74)。前記した通り、変数 T はベクトル量であり、ベクトルの各要素にトークン (オブジェクト) が代入される。

【 0 0 5 3 】

変数 i に 0 を代入して初期化し (ステップ 75)、 i が n より小さいかを判断する (ステップ 76)。ステップ 76 の判断が yes なら i が $n - 1$ であるかを判断する (ステップ 77)。つまり現在の i 番目のトークンが y 行における最後のトークンであるかを判断する。このステップ 77 の判断が no であるときには変数 t に i 番目のトークン $T[i]$ を代入する (ステップ 78)。その後変数 s にトークン t の右側に位置するスペースを代入する (ステップ 79)。そして、スペース s (トークン t の右側に位置する) がストリームに属するかを判断する (ステップ 80)。 $s.stream$ 関数はスペース s がストリームに属する時には新値

を戻す関数である。

【0054】

ステップ80の判断がyes（トークンtの右側のスペースがストリームである）の場合には、トークンtをテキストブロックデータベースに追加する（ステップ81）。なお、ステップ77でyes（トークンがy行における最後のトークンである）と判断された時にはステップ81に進む。

【0055】

一方ステップ80の判断がnoである時にはスペースsの大きさ|s|があらかじめ定めた最大スペース値(maxspace)より大きいかを判断し（ステップ82）、ステップ82の判断がyesの時にはステップ81に進んでトークンtをテキストブロックデータベースに追加する。ステップ82の判断がnoの時にはトークンt'としてi+1番目のトークンT[i+1]を代入し、さらにトークンtとt'とをマージしてトークンtとする。さらにトークンT[i]にトークンtを代入し、スペース列Sからスペースsを削除し、nから1を減じる（ステップ83）。つまり、ストリームでない空白の両側に位置するトークンをマージする処理を行う。その後ステップ76に進んで上記処理を繰り返す。

【0056】

なお、テキストブロックデータベースにトークンが記録された後は、iを1増加し（ステップ84）、ステップに進んで上記処理を繰り返す。

【0057】

そしてステップ76でnoと判断された時（1行分のトークンのマージ処理が終了した時）にはyを1増加し（ステップ85）、ステップ70に戻って処理を繰り返す。

【0058】

すなわち、上記処理により、ストリームまたは行の終端が検出されるまでは1行内のトークンはマージされる。このマージされたトークンが初期テキストブロックとして、テキストブロックデータベースに記録される。

【0059】

図14は、初期的テキストブロックが生成された段階の文書の例を示す図であ

る。前記処理フローより明らかな通り、この段階でのテキストブロックは1行内でのトークンのマージに止まるため、その深さは1である。また、同図に示すように、たとえば「Employees」と「in」との間の領域86のようにストリームの一部であるスペースが初期テキストブロックの間に残る。また、たとえば領域87のように、ストリームでないスペースによってもトークンがマージされない場合がある。つまり、ステップ82の判断において最大スペース値(maxspace)を越えた場合である。

【0060】

次に、初期リンクの生成を行う。初期リンクの生成は、テキストブロックデータベースとスペースデータベースを用いて計算する。図15は初期リンク生成の一例を示したフローチャートである。文書データの列方向の指標 y に0を代入して初期化し(ステップ88)、 y が最大行数 $max\ y$ より小さいかを判断する(ステップ89)。ステップ89の判断がnoなら処理を終了し(ステップ90)、yesなら以下の処理を行う。

【0061】

y 行に存在するテキストブロックを変数 T に代入し(ステップ91)、変数 $next$ に $y+1$ を代入する(ステップ92)。次に $next$ 行が空であるかを判断し(ステップ93)、空である場合には変数 $next$ をさらに1つ増加し(ステップ94)、空でない場合にはそのまま次のステップ95に進む。ステップ95では $next$ 行に存在するテキストブロックを変数 T' に代入する。そして変数 n に T の数 $|T|$ を代入する(ステップ96)。

【0062】

変数 i に0を代入して初期化し(ステップ97)、 i が n より小さいかを判断する(ステップ98)。ステップ98の判断がyesなら、変数 t に i 番目のテキストブロック $T[i]$ を代入した後(ステップ99)、 i が $n-1$ であるかを判断する(ステップ100)。つまり現在の i 番目のテキストブロックが y 行における最後のテキストブロックであるかを判断する。このステップ100の判断がnoであるときには、変数 t' に $i+1$ 番目のテキストブロック $T[i+1]$ を代入し(ステップ101)、 t と t' との間のスペースを s に代入する(ステ

ップ102)。sがストリームであるか(s.stream==true)もしくはsの大きさ|s|が最大リンクスペース(max link space)を超えているかを判断し(ステップ103)、ステップ103の判断がnoの時にはtとt'との間にリンクを生成する(ステップ104)。つまりステップ99からステップ104までの処理により、同一行内の隣接するテキストブロック間にストリームでないスペースが存在しかつそのスペースの大きさ(長さ)が最大リンクスペースより小さい時には隣接テキストブロック間にリンクを形成する。

【0063】

次に変数Lにtよりも左側に存在するT'内の全てのテキストブロックを代入する(ステップ105)。そして、tとLを構成する各々のテキストブロック間にリンクを生成する(ステップ106)。つまり、着目しているテキストブロックtの次行(次行が空行の時にはその次の行)に存在し、tよりも左側に位置するテキストブロックの全てにリンクを形成する。なお、ステップ100、103においてその判断がyes(着目テキストブロックtに右隣接するテキストブロックがない、あるいはストリームまたはスペースが大きくてリンクを張るのが妥当でない)と判断された時にはステップ105に進む。

【0064】

上記のようにリンクを形成した後iを1増加し(ステップ107)、ステップ98に進んで上記処理を繰り返す。そしてステップ98でnoと判断された時(1行分のテキストブロックについてリンク形成処理が終了した時)にはyを1増加し(ステップ108)、ステップ89に戻って処理を繰り返す。

【0065】

すなわち、上記処理により、初期テキストブロック間のリンクが形成される。このリンク生成の判断基準は、隣接テキストブロック間のスペースがストリームでないこと、スペースが予め定めた最大リンクスペースを越えないこと、リンクの終点となるテキストブロック(シンク)の位置が、リンクの始点となるテキストブロック(ソース)の次行左側に位置することである。このような条件を満たせば自動的に初期リンクが生成される。初期リンクはリンクのソース、シンクとなるテキストブロックの情報とともにデータベースに記録されるのは勿論である

。

【 0 0 6 6 】

以上のようにして初期化処理が終了する。図 1 6 は初期化処理のステップを擬似コードで表した図である。なお、擬似コードを表す図において左側に示した数字は行番号であり、以下同様である。

【 0 0 6 7 】

文書データdocをトークン化関数tokenizeに入力し、トークンtokensおよびスペースspacesを得る（行番号1）。また、文書データdocをストリーム関数steamに入力し、ストリームstreamsを得る（行番号2）。tokens、spaces、streamsを初期テキストブロック生成関数get_initial_blocksに入力し、初期テキストブロックとしてtext_blocksを得る（行番号3）。text_blocksを初期リンク生成関数get_initial_linksに入力し、初期リンクとしてlinksを得る（行番号4）。そして、初期テキストブロックtext_blocksおよび初期リンクlinksを文書グラフdocument_graphとしてストアする（行番号5）。文書グラフは、text_blocksをノード、linksを弧とするグラフセットである。

【 0 0 6 8 】

次に、単一要素サブクラスタの結合処理（ステップ32）を説明する。図17は単一要素サブクラスタ結合処理の一例を示した擬似コードを示す図であり、図18はそのフローチャートである。

【 0 0 6 9 】

まず、cluster関数を用いて文書をクラスタ化する（行番号1）。クラスタ化されたデータはcluster_setに格納される。文書のクラスタ化は前記した初期化処理で生成したテキストブロックとリンクで表現される文書からグラフを取り出すことにより行う。リンクで結合されているテキストブロックの集合が1つのグラフに対応する。

【 0 0 7 0 】

次にcluster_setに含まれるクラスタcの全てについてsub-cluster関数を用いてサブクラスタ化する。抽出されたサブクラスタはsub_cluster_setに格納される（行番号2, 3、ステップ109, 110）。サブクラスタ化は、クラスタか

ら単一要素のサブクラスタを抽出する作業である。たとえば各ノードにおける弧（リンク）の入次数および出次数が1となる条件（単一要素サブクラスタの定義）を満足するかをチェックしながら抽出できる。

【0071】

次にsub_cluster_setに含まれるサブクラスタsの全てについて、各サブクラスタsに含まれるリンクの妥当性を言語モデルを用いて評価する（行番号4～6、ステップ111～113）。リンクの妥当性は、リンクの始点（ソース）のテキストブロックとリンクの終点（シンク）のテキストブロックとが、言語モデルにおいて高い確率で出現する表現であるかを評価することにより行う。言語モデルにはたとえばNグラムモデルを用いることができる。ただしNグラムモデルに限らず、その他のモデルであっても構わない。評価が妥当である場合にはソースとシンクのテキストブロックがマージされる（行番号6、ステップ114）。

【0072】

次に、クラスタ結合処理（ステップ33）を説明する。図19はクラスタ結合処理の一例を示した擬似コードを示す図であり、図20はそのフローチャートである。クラスタの結合処理は複雑度の小さなものから順に行う。

【0073】

まず、複雑度に3を代入しておき（行番号1）、最大複雑度を取得する（行番号2）。次に複雑度が最大複雑度よりも小さいときには以下の処理を行う（行番号3）。

【0074】

cluster関数を用いて文書をクラスタ化し（行番号5）、cluster_setに含まれるクラスタcの全てについて、各クラスタcに含まれるリンクの複雑度をチェックする（行番号6～8、ステップ115～116）。リンクの複雑度が現ループの複雑度より小さければ（行番号8、ステップ116）、リンクがユニークに妥当であると評価できるかを判断する（行番号9、ステップ117）。判断の結果yesであればリンクのソースとなるテキストブロックとシンクとなるテキストブロックをマージする（行番号9、ステップ118）。なおリンクがユニークに妥当であるかどうかとは、唯一確かなリンクしか存在しないことをいう。たとえ

ば「number」「of」間のリンクと「of」「of」間のリンクが並存する場合、言語モデルからは「of」「of」間のリンクはありえない。この場合、「of」「of」間のリンクが取り去られ、「number」「of」間のリンクが選択される。そして「number」「of」がマージされて新たなテキストブロック「number of」が生成される。

【0075】

次に、クラスタ間の接続評価ステップ（ステップ34）を説明する。図21はクラスタ間接続評価処理の一例を示した擬似コードを示す図であり、図22はそのフローチャートである。この処理はクラスタの結合処理（ステップ33）と類似している。ステップ33では複数のリンクを評価した時に、唯一選択し得るリンクが存在する時にはこのリンクを選択する処理を行った。ここでは、複数のリンクが存在し、何れも選択可能な時の処理を説明する。なお、ステップ33と同様な事項の説明は省略する。

【0076】

ステップ34では、cluster_setに含まれるクラスタcの全てについて、各クラスタcに含まれるリンクの順位付けを行う（行番号7、ステップ119）。順位付けられたリンクordered_linksに含まれる各リンクについて（行番号8、ステップ120）、リンクの複雑度をチェックし（行番号9、ステップ121）、リンクの妥当性評価に有意な差があるかを判断する（行番号10、ステップ122）。リンクの妥当性評価には、たとえば言語モデルによる出現確率を用いることができる。出現確率が高いほど接続妥当性は高くなる。なおその他の妥当性評価および有意差の判定手法を用いることができることは勿論である。前記判断の結果yesであればリンクのソースとなるテキストブロックとシンクとなるテキストブロックをマージする（行番号10、ステップ123）。なお、有意差が見られない時にはリンクは並存したままである。

【0077】

このようにして空白でレイアウトされた文書から意味のあるテキストブロックを自動的に生成できる。図23は最終的な処理後の出力の一例を示した図である。言語モデルにより妥当と判断された接続をマージして得られたテキストを含む

テキストブロックが表示されている。図示するように本実施の形態の処理によるテキストブロックは原文書の空間座標情報を維持したまま出力される。つまり最終的なテキストブロックの位置は、文書の位置座標で特定できる。たとえばテキストブロックの左上の座標と右下の座標で位置が特定される。このように原文書の空間情報が維持され、かつ、各テキストブロック内のテキストは言語モデルで保証された意味のある内容を含む。このため、本実施の形態の処理方法を前処理手段として高度な自然言語処理（たとえばテキストマイニングや機械翻訳）を原文書の内容を欠落することなく適用できる。さらに、本実施の形態の文書処理方法は、空間的な位置情報とスペースを利用してテキストブロックを生成するので、複雑なレイアウトを持つ文書にも容易に適用することが可能である。このため、適用の対象となる文書の範囲を広げ、より汎用的な文書処理に供することができる。

【 0 0 7 8 】

以上、本発明者によってなされた発明を発明の実施の形態に基づき具体的に説明したが、本発明は前記実施の形態に限定されるものではなく、その要旨を逸脱しない範囲で種々変更可能である。

【 0 0 7 9 】

たとえば、入力文書には、前記実施の形態で用いた表に限らず、2 段組等の複数段表示文書、リスト、表題が付された文書、ダブルスペース文書、マルチカラムセルあるいはマルチロウセルを有する表文書、サブセルを有する表文書、省略形のリストあるいは表等の文書にも適用できる。適用可能な文書の例を図 2 4 の示す。

【 0 0 8 0 】

また、日本語等縦書きを含む文書にも適用できる。前記実施の形態では横書き文書を前提に説明したが、縦書き文書に適用するようにアルゴリズムを変更することは容易である。また、横書きと縦書きが混在するような文書にも適用できる。

【 0 0 8 1 】

さらに 1 つのテキストが複数のテキストにかかる場合や、複数のテキストが 1

つのテキストにかかる場合のように文書間のかかり受けが複雑になる場合にも本実施の形態を用いれば正確に把握できる。この場合1つのオブジェクト（テキストブロック）に複数のリンクが入る（または出る）状態で記録される。

【0082】

また、前記実施の形態では、単一要素のサブクラスタにおけるテキストブロック結合処理（ステップ32）、クラスタ結合処理（ステップ33）、およびクラスタ間接続評価（ステップ34）の全てのステップを有する例を説明したが、ステップ32あるいはステップ34については必須のステップではない。入力文書によっては単一要素サブクラスタが存在しない場合があり、また、唯一のリンク可能性が発見できない場合もある。この場合、ステップ34のテキストブロックの接続評価により、本実施の形態と同じ文書出力を得ることができる。ただし、ステップ32、33を入れることにより処理の効率化が図れる。

【0083】

【発明の効果】

本願で開示される発明により、表、箇条書き、多段組等任意にレイアウトされた文書から意味のあるテキストブロックを抽出することができる。

【図面の簡単な説明】

【図1】

（a）はスタンドアロンのコンピュータを構成の概略を示し、（b）は、コンピュータネットワークの場合を示す。

【図2】

本システムに入力される文書の一例を示す図である。

【図3】

キャラクタ、スペース、トークンの例を示す図である。

【図4】

テキストブロックの例を示す図である。

【図5】

(a)はグラフの一例を示し、(b)はにグラフセットの一例を示す図である。

【図6】

単一要素サブクラスタの一例を示す図である。

【図 7】

ストリームの一例を示す図である。

【図 8】

本発明の一実施の形態である処理の概要を示したフローチャートである。

【図 9】

トークン生成ステップの一例を示したフローチャートである。

【図 1 0】

トークン化処理が終了した後の文書の一例を示す図である。

【図 1 1】

ストリームサイズの計算方法の一例を示したフローチャートである。

【図 1 2】

ストリームを計算した後の結果の一例を示す図である。

【図 1 3】

初期テキストブロックの計算方法の一例を示したフローチャートである。

【図 1 4】

初期的テキストブロックが生成された段階の文書の一例を示す図である。

【図 1 5】

初期リンク生成の一例を示したフローチャートである。

【図 1 6】

初期化処理のステップを擬似コードで表した図である。

【図 1 7】

単一要素サブクラスタ結合処理の一例を示した擬似コードを示す図である。

【図 1 8】

単一要素サブクラスタ結合処理の一例を示したフローチャートである。

【図 1 9】

クラスタ結合処理の一例を示した擬似コードを示す図である。

【図 2 0】

クラスタ結合処理の一例を示したフローチャートである。

【図 2 1】

クラスタ間接続評価処理の一例を示した擬似コードを示す図である。

【図 2 2】

クラスタ間接続評価処理の一例を示したフローチャートである。

【図 2 3】

最終的な処理後の出力の一例を示した図である。

【図 2 4】

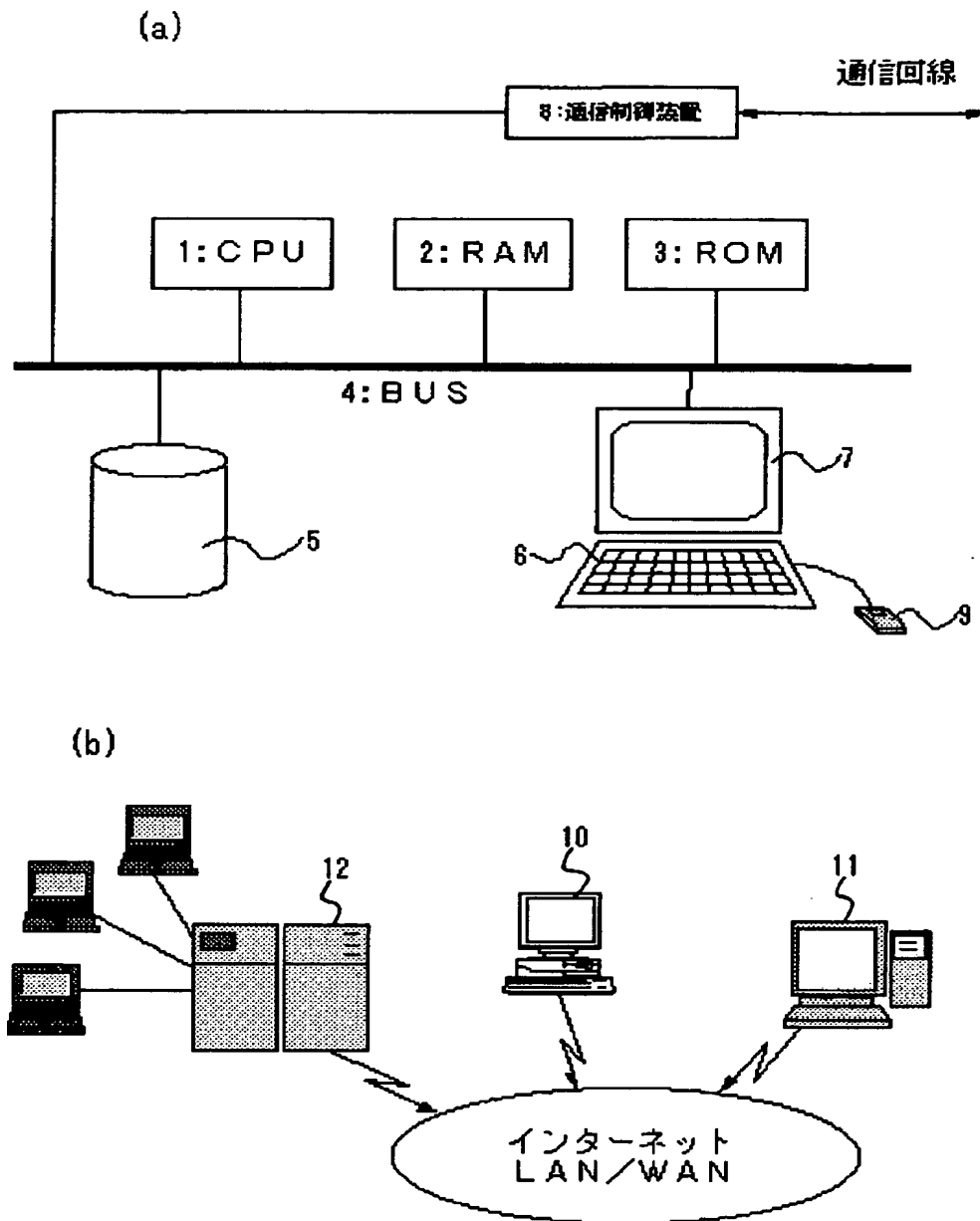
入力文書のその他の例を示す表図である。

【符号の説明】

1 … 中央演算処理装置、2 … 主記憶装置、3 … 不揮発性記憶装置、4 … バス、
5 … 外部記憶装置、6 … データ入力デバイス、7 … 表示デバイス（表示装置）、
8 … 通信制御装置、9 … マウス、10 … パーソナルコンピュータ、11 … ワーク
ステーション、12 … メインフレームコンピュータ、25, 27, 29 … ノード
、26 … 弧、T, t … テキストブロックまたはトークン、c … クラスタ、max
x … 最大文字数、max y … 最大行数、s … サブクラスタまたはスペース。

【書類名】 図面

【図 1】



【図 2】

0123456.....

	Individual Grants		Potential Realizable			
	Number of	Percent of	Value at Assumed			
	Securities	Total Options	Annual Rates of Stock			
	Underlying	Granted to	Price Appreciation			
	Options	Employees in	for Option Term			
	Granted (\$)	Fiscal Year	5 % (\$)	10% (\$)		
	Name	Expiration	Date			
0	Steven H. Rothman	31.3%	\$4.43	8/31/2001	\$0	\$30,000
1						
2	Howard Paveny	31.3%	\$4.43	8/31/2001	\$0	\$30,000
3						
4	Robert Fries	-	-	-	\$0	\$0
5						
6	Ramon Mota	3.1%	\$2.50	11/30/2001	\$7,450	\$12,650

maxy

maxx

maxy

【図 3】

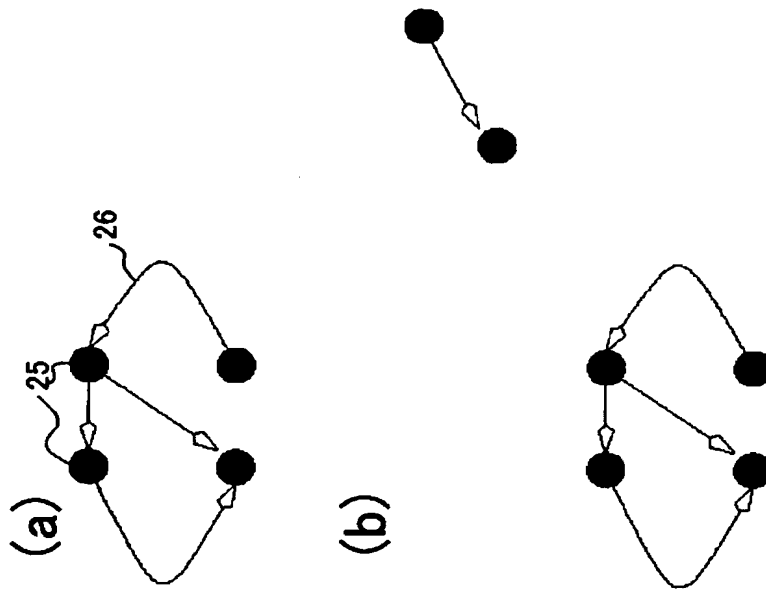
	Individual Grants	Number of Percent of Securities Total Options	Underlying Granted to Employees in Fiscal Year	Options Granted (#)	Price (\$/sh)	Expiration Date	for Option Form		
							5 % (\$)	10% (\$)	Potential Realizable Value at Assumed Annual Rates of Stock Price Appreciation
21	me			50,000	\$4.43	8/31/2001	\$0	\$30,000	
21	Steven H. Rothman			50,000	\$4.43	8/31/2001	\$0	\$30,000	
21	Bward Pavony			50,000	\$4.43	8/31/2001	\$0	\$30,000	
21	Robert Fries			-	-	-	\$0	\$0	
21	Samon Mota			5,000	\$2.50	11/30/2001	\$7,450	\$12,650	

【図4】

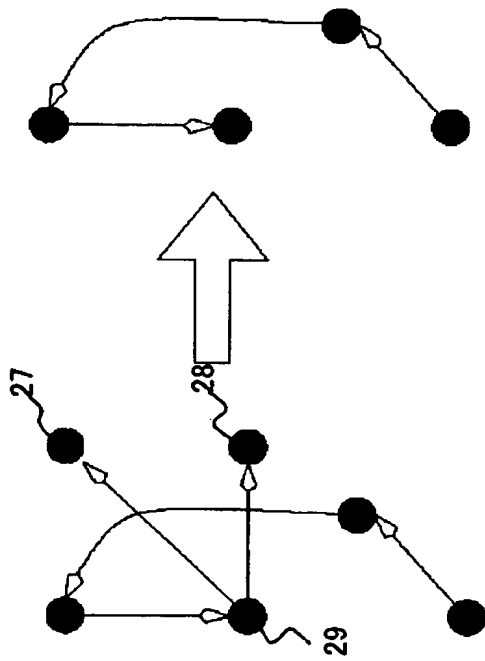
24

Name	Individual Grants				Potential Realizable		
	Number of	Percent of	Underlying	Options	Exercise	Expiration	Value at Assumed
	Shares	Total Options					Annual Rates of Stock
	Granted	Granted to	Employees in	Fiscal Year	Price (\$/sh)	Date	Price Appreciation
							for Option Term
							5 % (\$)
							10% (\$)
Steven H. Rothman	50,000	31.3%			\$4.43	8/31/2001	\$0 \$30,000
Howard Pavony	50,000	31.3%			\$4.43	8/31/2001	\$0 \$30,000
Robert Fries	-	-			-	-	\$0 \$0
Ramon Mota	5,000	3.1%			\$2.50	11/30/2001	\$7,450 \$12,650

【図 5】



【図 6】

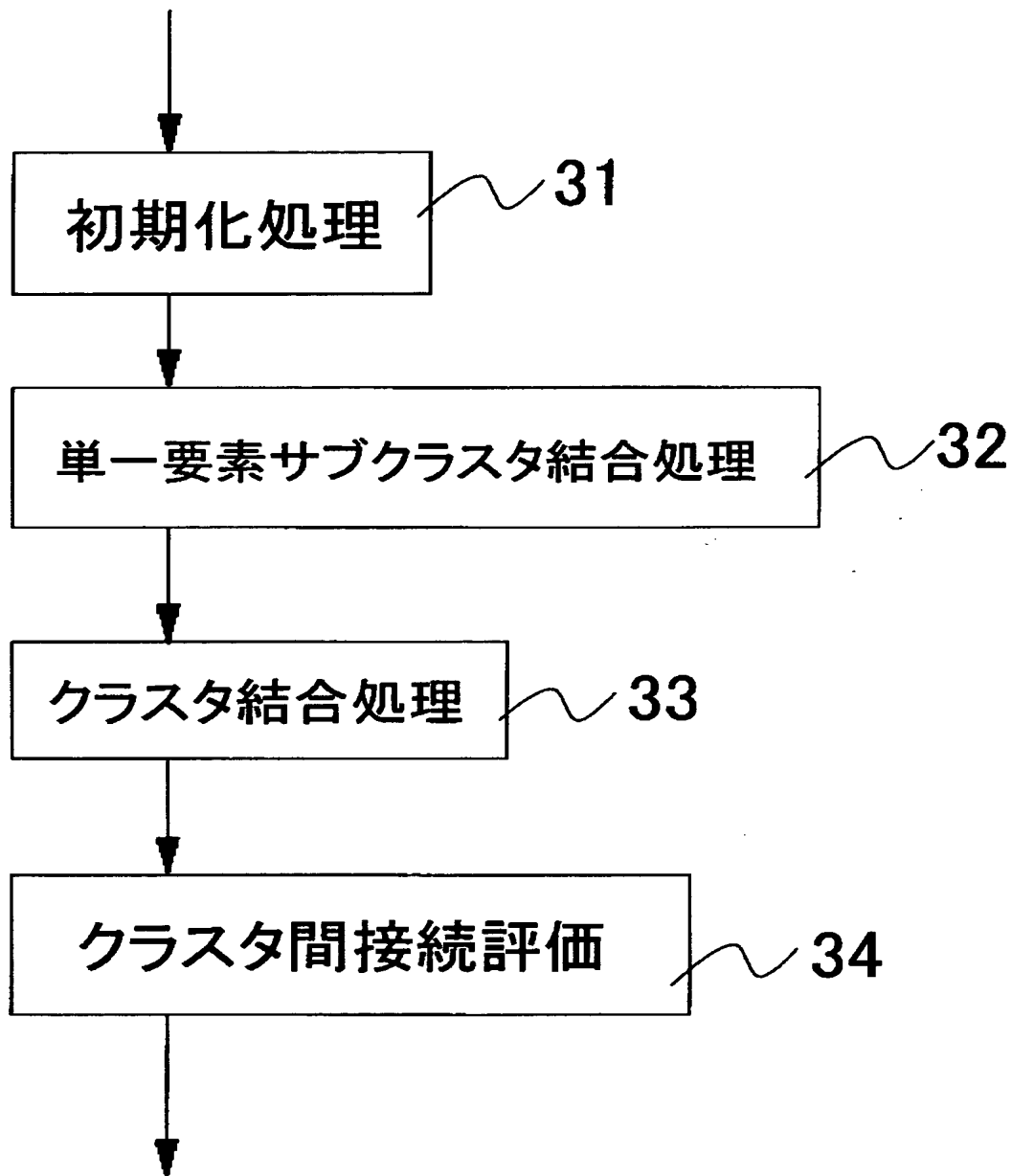


【図7】

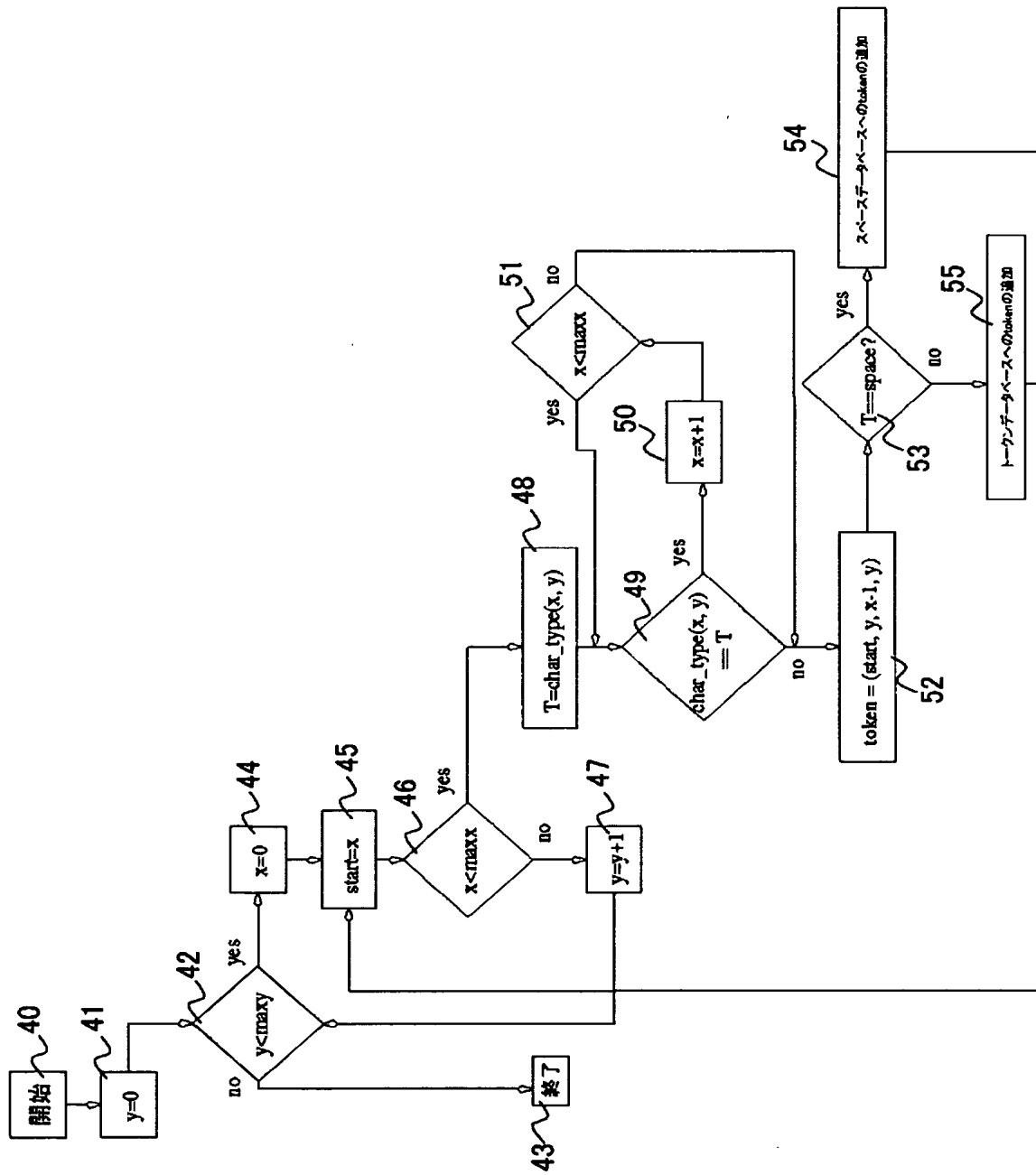
30

Individual Grants									
Number of Securities Granted		Percent of Total Options		Underlying Options Granted to Employees in Fiscal Year		Exercise Price (\$/sh)		Expiration Date	
Name	Granted (#)	Fiscal Year	Price (\$/sh)	Options Granted	Percent of Total	Exercise Price (\$/sh)	Expiration Date	Potential Realizable Value at Assumed Annual Rates of Stock Price Appreciation for Option Term 5 % (\$)	10% (\$)
Steven H. Rothman	50,000	31.3%	\$4.43	8/31/2001	\$0	\$30,000			
Howard Pavony	50,000	31.3%	\$4.43	8/31/2001	\$0	\$30,000			
Robert Fries	-	-	-	-	\$0	\$0			
Ramon Mota	5,000	3.1%	\$2.50	11/30/2001	\$7,450	\$12,650			

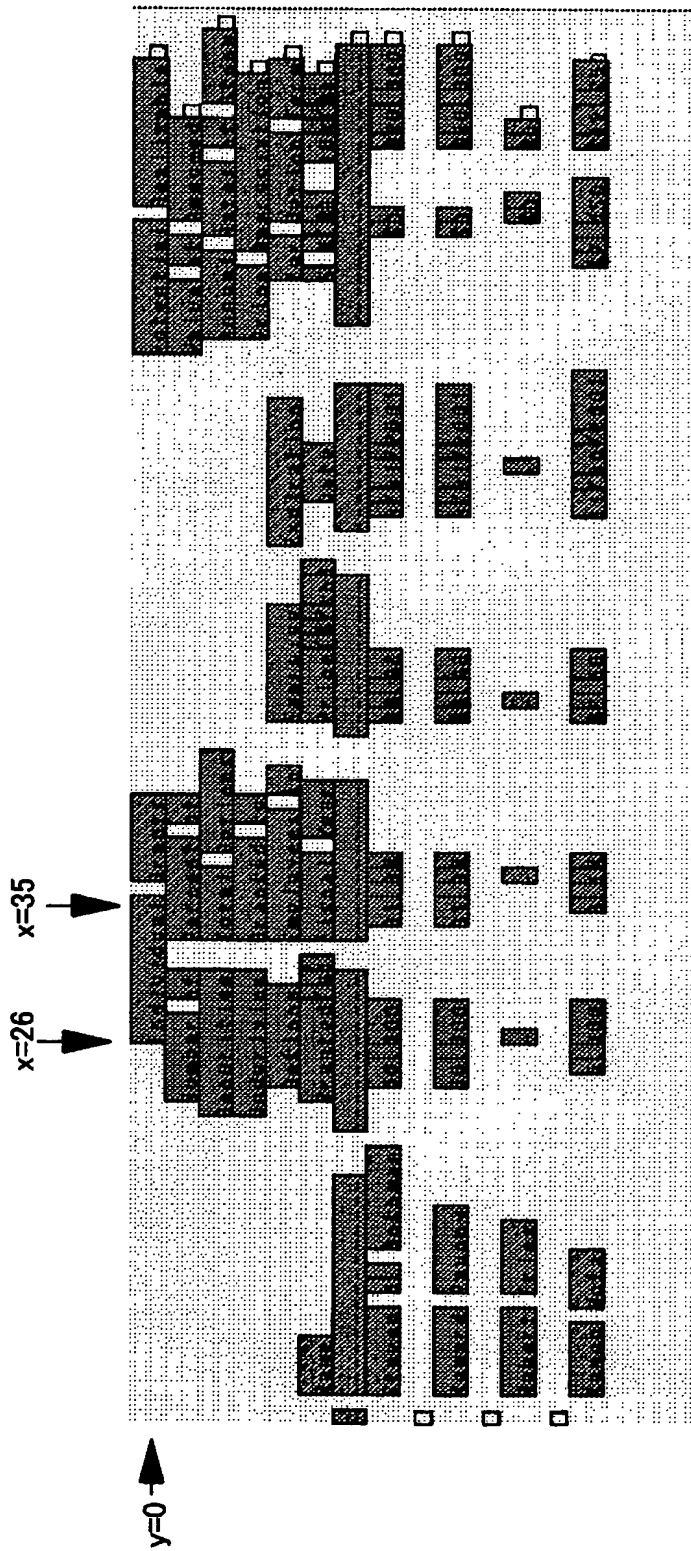
【図 8】



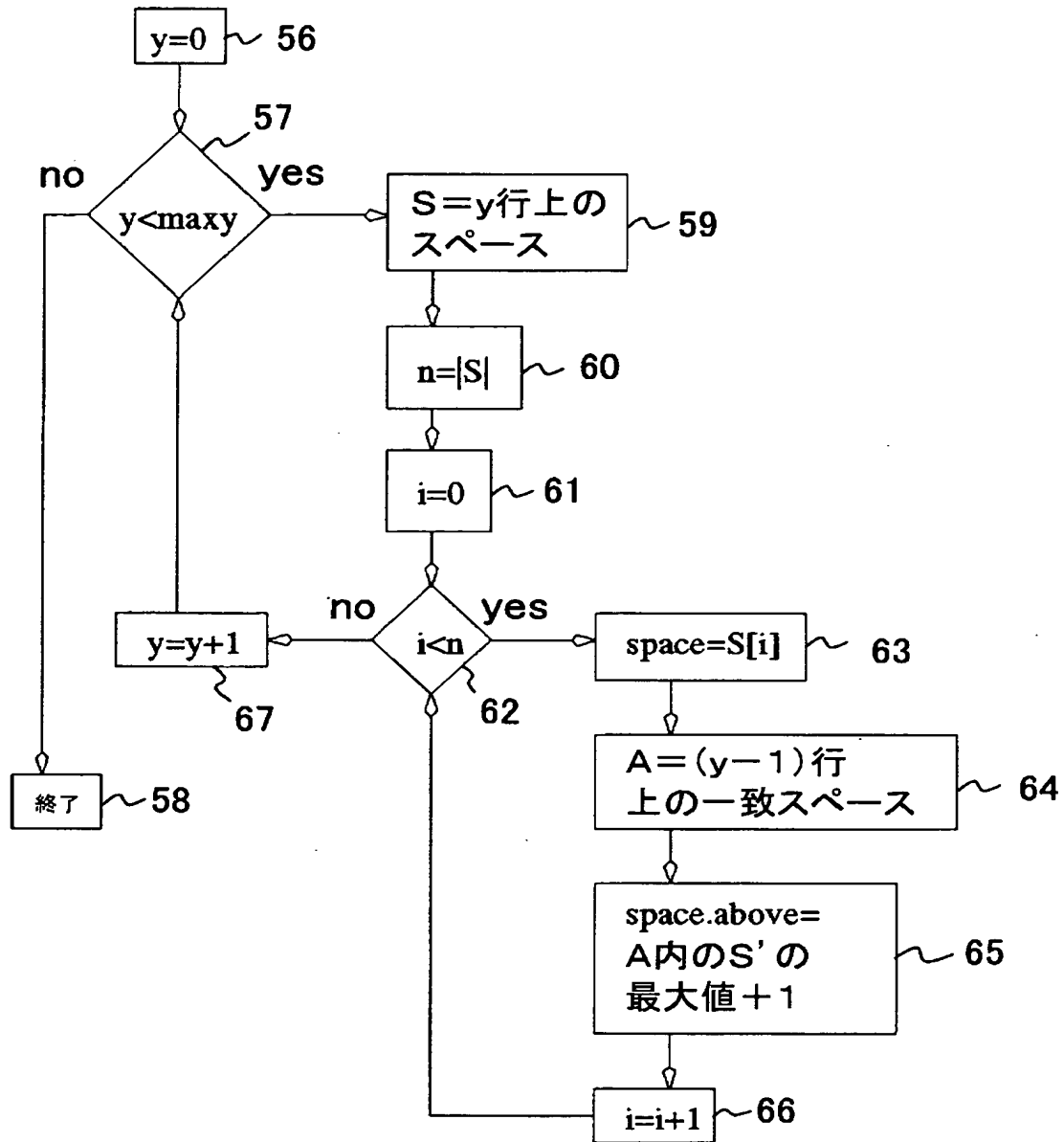
【図 9】



【図 10】



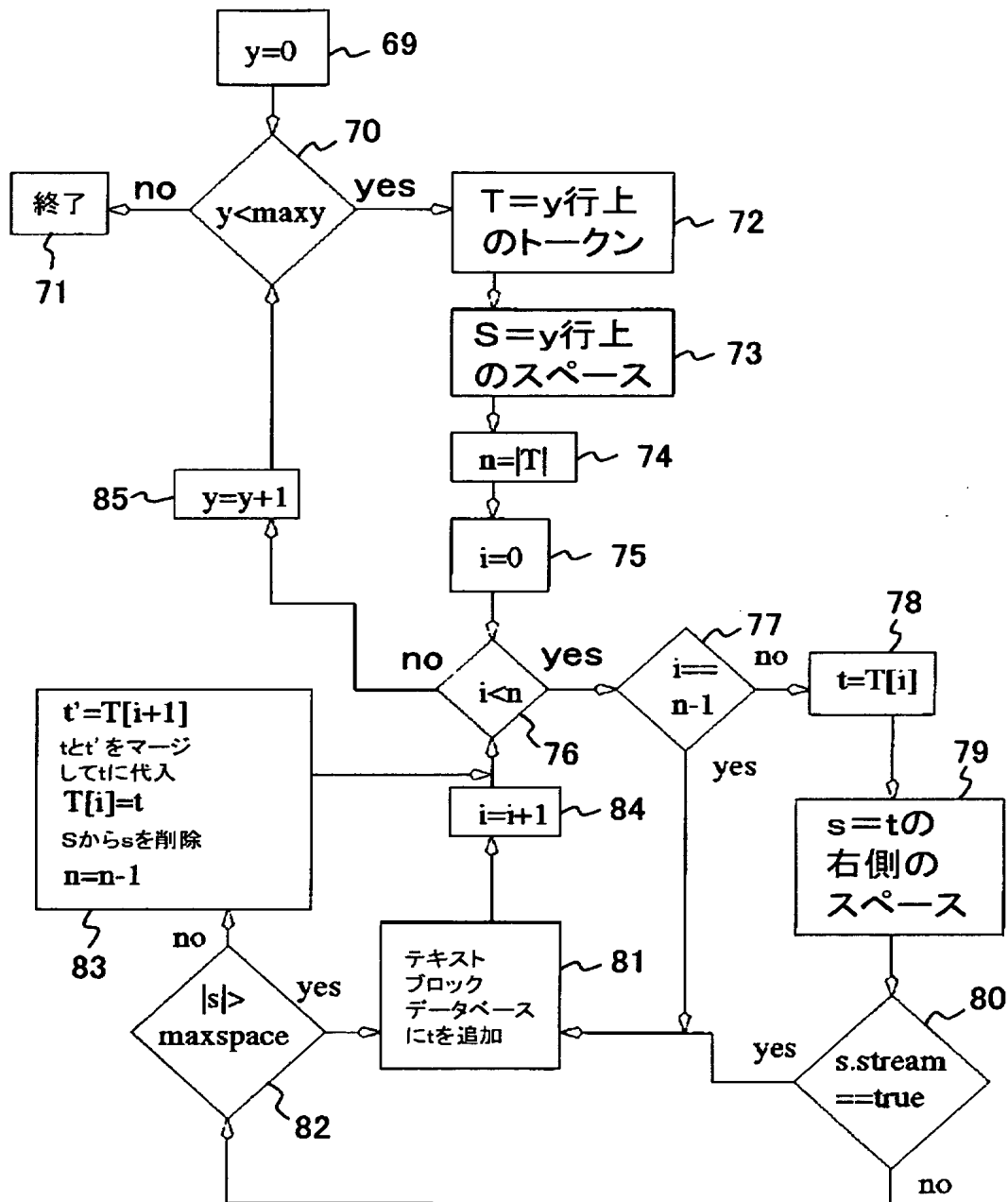
【図 11】



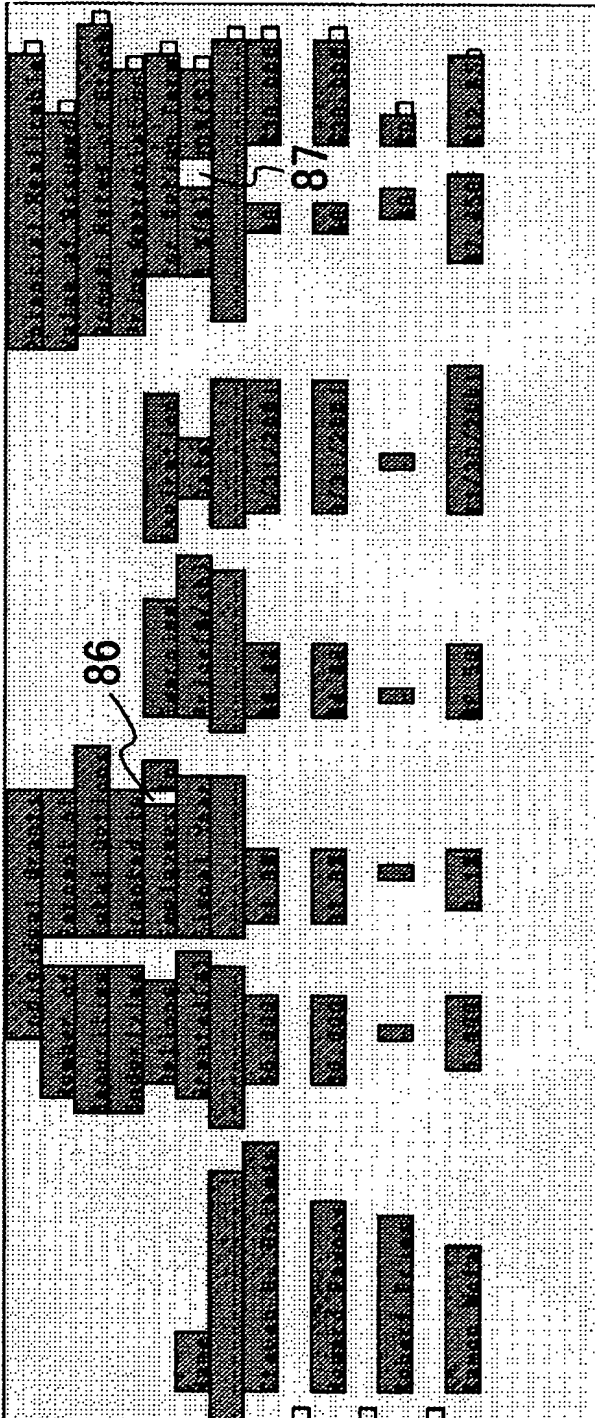
【图 1 2】

Name	Options Granted (#)	Fiscal Year	Exercise Price (\$/sh)	Expiration Date	For Option Term	10% (\$)	Potential Realizable Value at Assumed Annual Rates of Stock Price Appreciation
Steven H. Rothman	50,000	91.3%	\$4.43	8/31/2001		\$0	\$30,000
Howard Pavony	50,000	91.3%	\$4.43	8/31/2001		\$0	\$30,000
Robert Fries	-	-	-	-		\$0	\$0
Ramon Mota	5,000	3.1%	\$2.50	11/30/2001		\$7,450	\$12,650

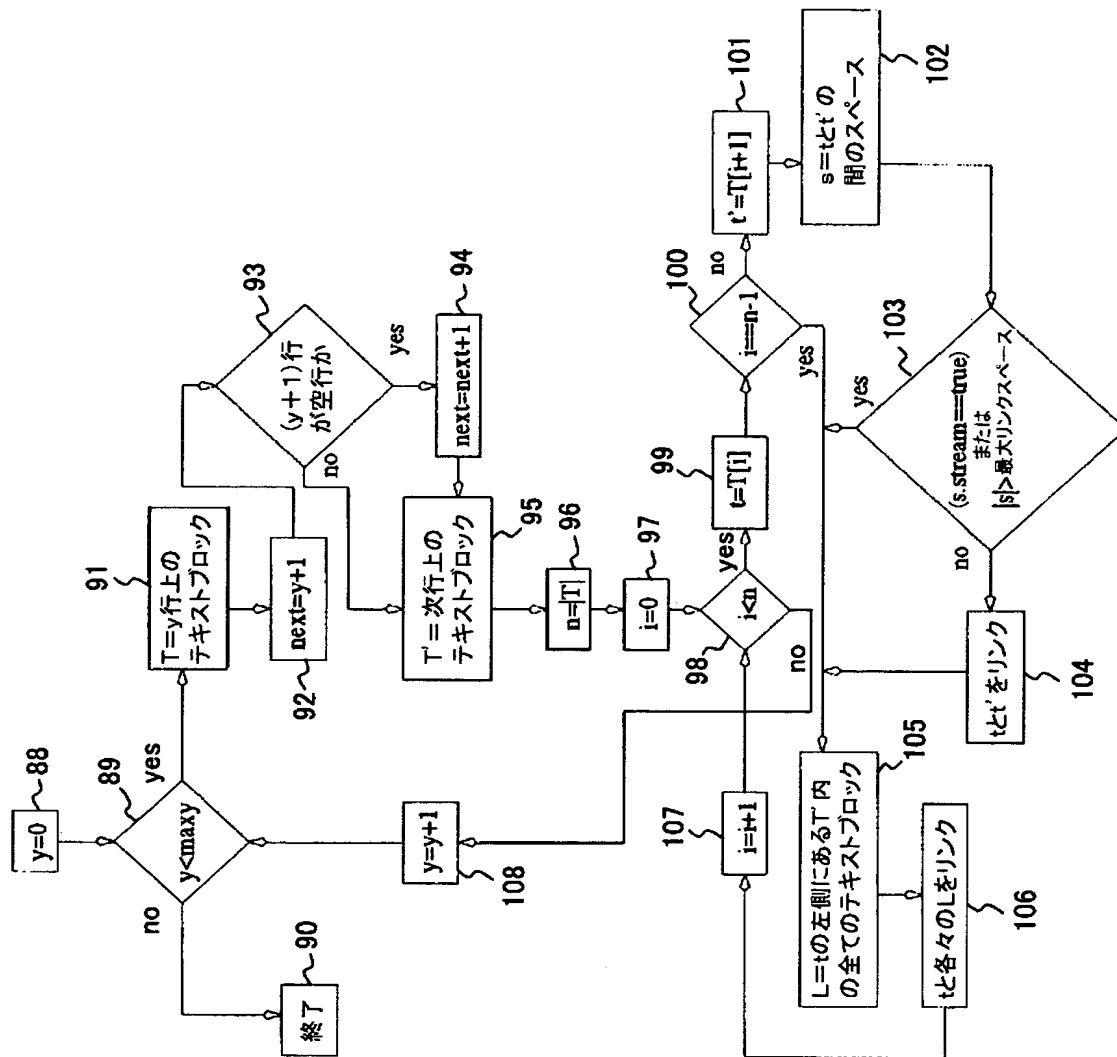
【図 13】



【図 1 4】



【図 15】



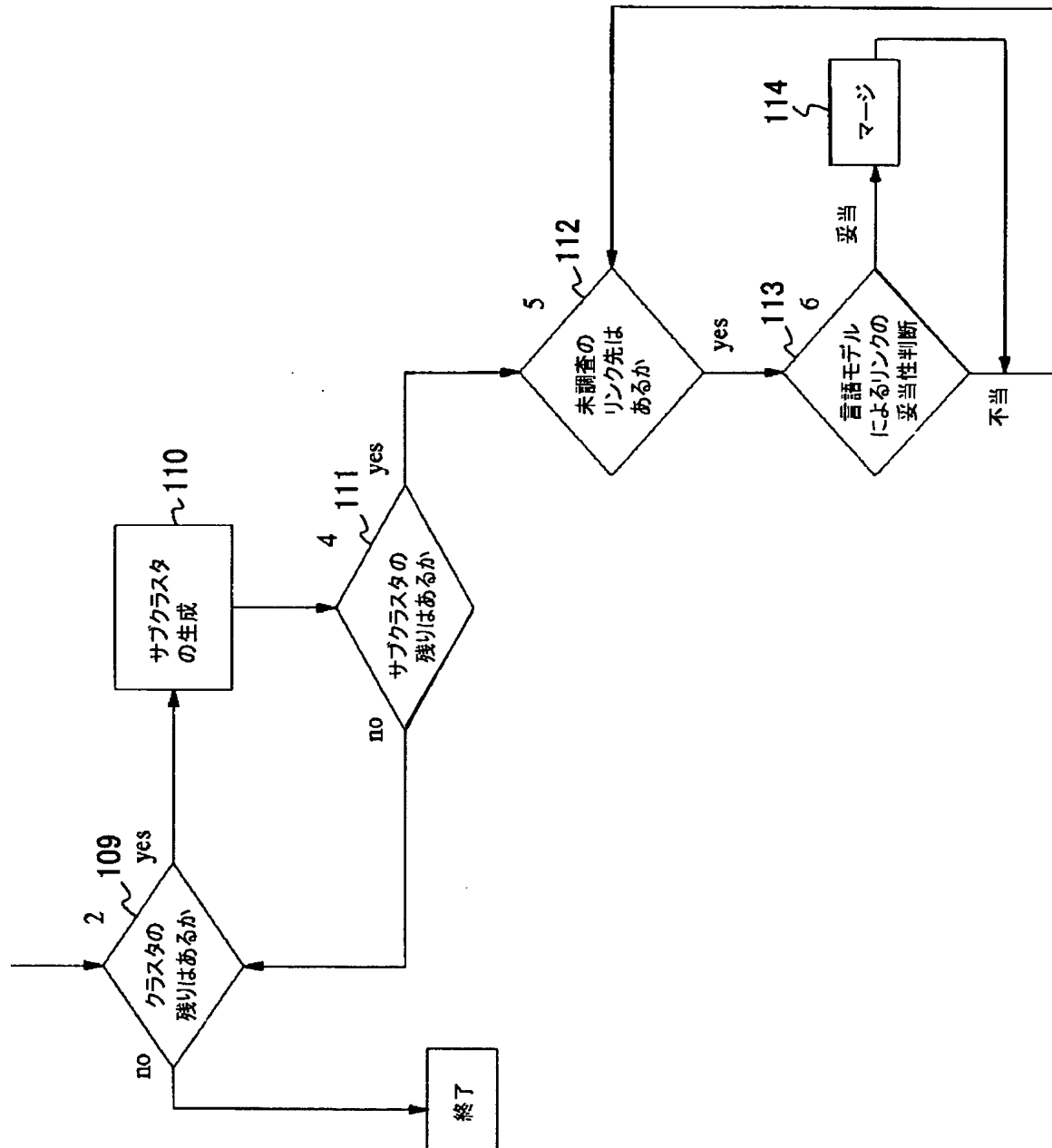
【図 1 6】

```
1 (tokens, spaces)←tokenize(doc);  
2 streams←stream(doc);  
3 text_blocks←get_initial_blocks(tokens, spaces, streams);  
4 links←get_initial_links(text_blocks);  
5 document_graph←(text_blocks, links);
```

【図 1 7】

```
1 cluster_set←cluster(doc);
2 for all c∈ cluster_set do {
3   sub_cluster_set←sub-cluster(c);
4   for all s∈ sub_cluster_set do {
5     for all links in s do {
6       if valid(link) then merge(sink, source);
7     }
8   }
9 }
```

【図 18】



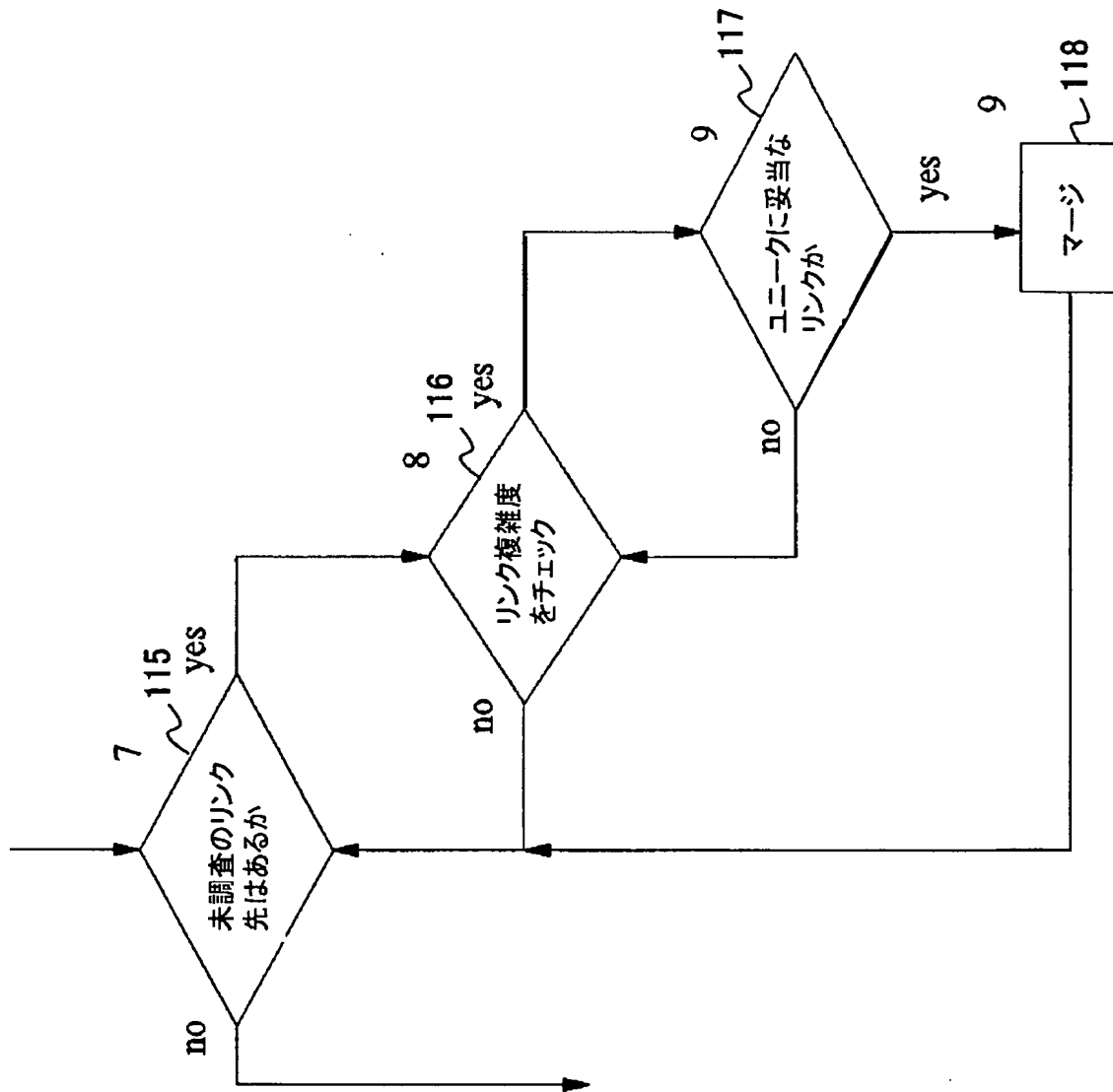
【図 1 9】

```

1 perplexity←3;
2 max_perplexity←get_max_perplexity;
3 while(perplexity<max_perplexity) do {
4   repeat while merges continue to be carried out
5   cluster_set←cluster(doc);
6   for all c∈ cluster_set do {
7     for all links in c do {
8       if(perplexity(link)<perplexity) then do {
9         if unique_valid_link(link) then merge(sink, source);
10      }
11    }
12  }
13  perplexity←perplexity + 1;
14  max_perplexity←get_max_perplexity;
15}

```

【図 20】



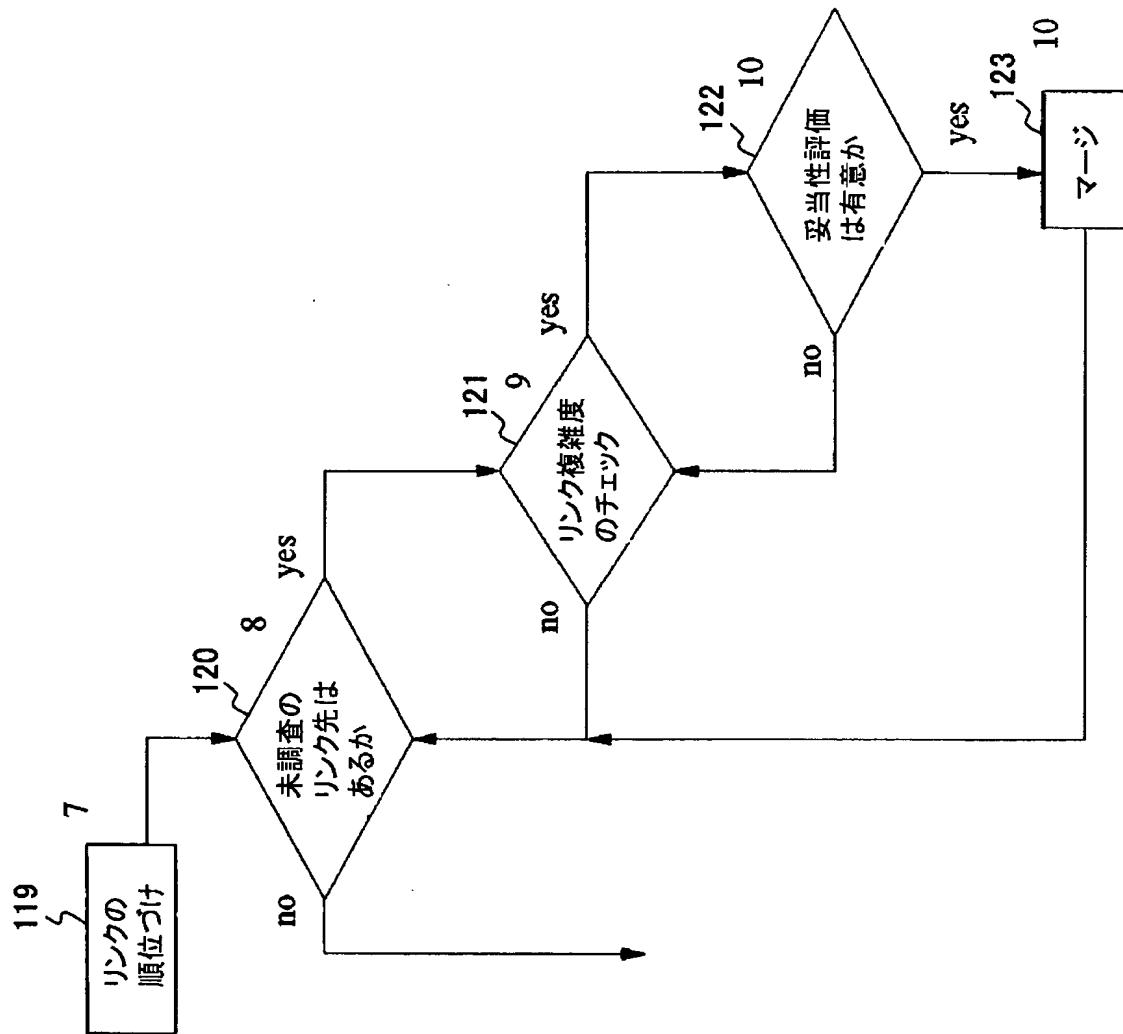
【図 2 1】

```

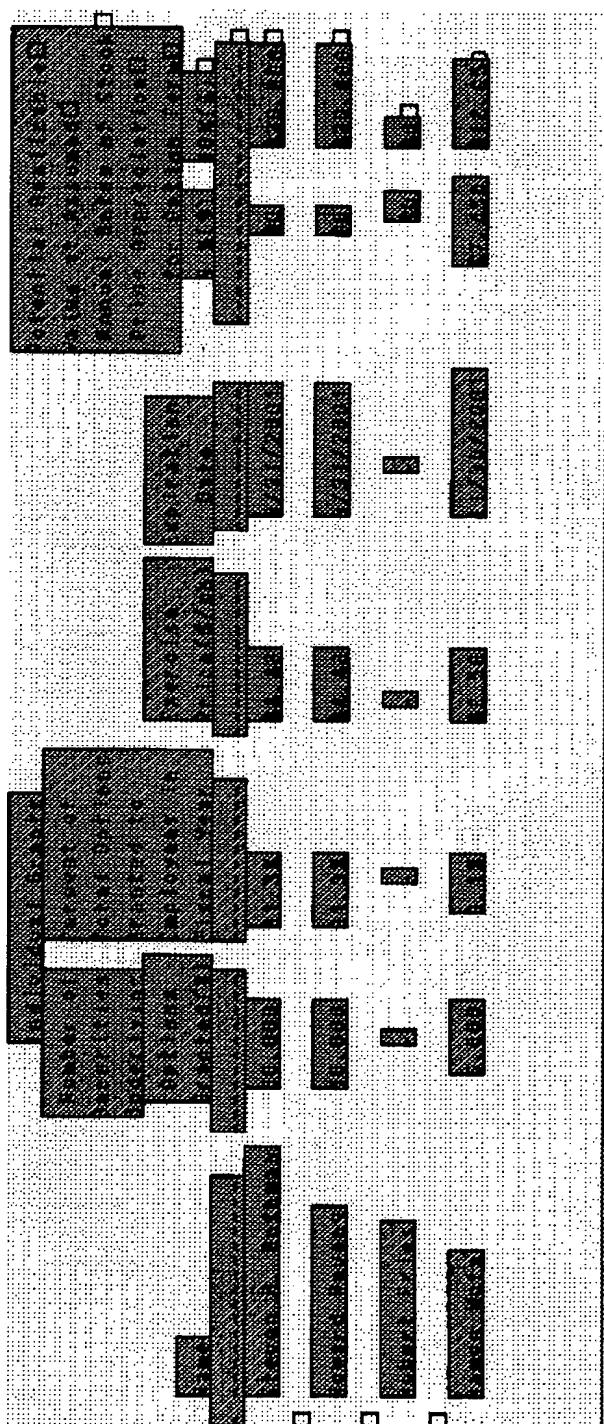
1 perplexity←3;
2 max_perplexity←get_max_perplexity;
3 while(perplexity<max_perplexity) do {
4     repeat while merges continue to be carried out
5     cluster_set←cluster(doc);
6     for all c∈ cluster_set do {
7         ordered_links←get_ordered_links(c);
8         for each link ∈ ordered_links do {
9             if perplexity(link)<perplexity then do {
10                if distinguished_valid_link(link) then merge(sink, source);
11            }
12        }
13    }
14    merge_unary_sub_clusters;
15    perplexity←perplexity + 1;
16    max_perplexity←get_max_perplexity;
17}

```

【図 22】



【图 2 3】



【图 2 4】

Double Column	For example, a paragraph splitting this at best is geometrically objectionable to the casual reader; the line organization of a double column of text will indicate where the line breaks occur.	Multi-column cell	Page Date Number of Days Cells Boxes
False White Space Rivers	Blank lines between any complete or false fold lines of rivers or white space which look like unprinted blank lines but in fact contain text.	Multi-row Cell	Page Date Number of Days Cells Boxes
Approved/Marginal Material	APPROVED MARGINAL The entire of such-story items collected or on signed and accepted or the lower for attention in the firm	Multi-row Stub Cells	Page Date Number of Days Cells Boxes
Simple Ap- proved/Marginal Material	STANDARD TO STYLE: If the sentence falls to meet on 1875 dollar order with	Multiple Cells (below header)	Page Date Number of Days Cells Boxes
Unmarked Headings	Adjustable spot changes in column width	Elliptical Cell Contents	Page Date Number of Days Cells Boxes
Double Spacing	In the event that the number of extending above of some part of the company is changed by reason of recognition, identification, and	Grid Quantization	Page Date Number of Days Cells Boxes
Elliptical Lists	At the underlined direction, stand to the instructions of this report of condition and declare that it has been modified by me, and to the best of our knowledge and	Orientation Variation	Page Date Number of Days Cells Boxes
Short Paragraphs	In my cell the immediate: -- through subordinates, -- through agents or -- directly to a limited number of horizontal perimeters or to a single partner.	One to Many, Many to One	Page Date Number of Days Cells Boxes

【書類名】 要約書

【要約】

【課題】 表、箇条書き、多段組等任意にレイアウトされた文書から意味のあるテキストブロックを抽出する。

【解決手段】 空白等でレイアウトされた文書を入力し、文書の空間座標で関連付けたシンボルを取得する。シンボルから同一タイプのキャラクタの連続を抽出しトークンとスペースを生成する。列方向に連続したスペースからストリームを生成し、ストリームとトークンからテキストブロックを生成する。テキストブロック間のリンクを生成して、文書グラフとする。文書グラフ内のテキストブロック間の接続（リンク）の妥当性を言語モデルを用いて評価し、接続が妥当な場合はそのテキストブロックをマージする。

【選択図】 図 8

認定・付加情報

特許出願の番号	特願 2 0 0 0 - 1 9 0 3 3 5
受付番号	5 0 0 0 0 7 9 2 2 6 1
書類名	特許願
担当官	塩崎 博子 1 6 0 6
作成日	平成 1 2 年 8 月 4 日

<認定情報・付加情報>

【特許出願人】

【識別番号】	390009531
【住所又は居所】	アメリカ合衆国 1 0 5 0 4、ニューヨーク州 アーモンク (番地なし)
【氏名又は名称】	インターナショナル・ビジネス・マシーンズ・コーポレーション

【代理人】

【識別番号】	100086243
【住所又は居所】	神奈川県大和市下鶴間 1 6 2 3 番地 1 4 日本アイ・ビー・エム株式会社 大和事業所内
【氏名又は名称】	坂口 博

【復代理人】

【識別番号】	100112520
【住所又は居所】	神奈川県大和市中心林間 3 丁目 4 番 4 号 サクライビル 4 階 間山・林合同技術特許事務所
【氏名又は名称】	林 茂則

【選任した代理人】

【識別番号】	100091568
【住所又は居所】	神奈川県大和市下鶴間 1 6 2 3 番地 1 4 日本アイ・ビー・エム株式会社 大和事業所内
【氏名又は名称】	市位 嘉宏

【選任した復代理人】

【識別番号】	100110607
【住所又は居所】	神奈川県大和市中心林間 3 丁目 4 番 4 号 サクライビル 4 階 間山・林合同技術特許事務所
【氏名又は名称】	間山 進也

【選任した復代理人】

【識別番号】	100098121
--------	-----------

認定・付加情報（続き）

【住所又は居所】 神奈川県大和市中央林間3丁目4番4号 サクラ
イビル4階 間山・林合同技術特許事務所
【氏名又は名称】 間山 世津子

出 願 人 履 歴 情 報

識別番号 [390009531]

1. 変更年月日 2000年 5月16日

[変更理由] 名称変更

住 所 アメリカ合衆国10504、ニューヨーク州 アーモンク (番地なし)

氏 名 インターナショナル・ビジネス・マシーンズ・コーポレーション

**This Page is Inserted by IFW Indexing and Scanning
Operations and is not part of the Official Record**

BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images include but are not limited to the items checked:

- ☐ BLACK BORDERS
- ☐ IMAGE CUT OFF AT TOP, BOTTOM OR SIDES
- ☐ FADED TEXT OR DRAWING
- ☒ BLURRED OR ILLEGIBLE TEXT OR DRAWING
- ☐ SKEWED/SLANTED IMAGES
- ☐ COLOR OR BLACK AND WHITE PHOTOGRAPHS
- ☐ GRAY SCALE DOCUMENTS
- ☐ LINES OR MARKS ON ORIGINAL DOCUMENT
- ☐ REFERENCE(S) OR EXHIBIT(S) SUBMITTED ARE POOR QUALITY
- ☐ OTHER: _____

IMAGES ARE BEST AVAILABLE COPY.

As rescanning these documents will not correct the image problems checked, please do not report these problems to the IFW Image Problem Mailbox.